O‘ZBEKISTON RESPUBLIKASI OLIY VA O‘RTA MAXSUS TA’LIM VAZIRLIGI
ALISHER NAVOIY NOMIDAGI TOSHKENT DAVLAT O‘ZBEK TILI VA ADABIYOTI UNIVERSITETI

O‘ZBEK FILOLOGIYASI FAKULTETI

“Himoyaga tavsiya etilsin”
Fakultet dekani
____S.Abdullayev
“___ “ _____2018-yil

5120100 – Filologiya va tillarni o‘qitish (o‘zbek tili) ta’lim yo‘nalishi IV kurs 403-guruh talabasi Nurullayeva Oydin Nazirjon qizining “O‘zbek tilining milliy korpusini yaratish tamoyillari” mavzusida yozilgan

BITIRUV MALAKAVIY ISHI

Ilmiy rahbar:
_____ f.f.d., B.Mengliyev,
O‘zbek tilshunosligi kafedrasi professori

Taqrizchi:

_____f.f.d., Y.Odilov,
O‘zRFA Til, adabiyot va folklor instituti
katta ilmiy xodimi

“Himoyaga tavsiya etilsin”
O‘zbek tilshunosligi kafedrasi mudiri,
f.f.d., prof., _____ M.Qurbonova
“_____“ _____2018-yil.

TOSHKENT – 2018

MUNDARIJA

KIRISH

Mavzuning dolzarbligi. Til ma’naviy merosimizni asrash va boyitishning asosiy vositasi hisoblanadi. Tillarni saqlab qolish milliy ma’naviyatni saqlab qolish demakdir. Zero, asl ma`naviyat faqat milliy shakldagina mavjud bo‘ladi. Mamlakatimizda tilga e`tibor ma’naviyatga e`tiborning ustuvor yo‘nalishlaridan biri darajasiga ko‘tarildi. Shu bois ona tilimizni avaylab-asrash, boyitish, undan amaliy foydalanish samaradorligini oshirish bilan birga, o‘zbek tilining zamonaviy axborot-kommunikatsiya tizimida keng qo‘llanishiga erishish kechiktirib bo‘lmaydigan dolzarb vazifaga aylandi. Chunki ona tilimizning jahonga chiqishiga erishish milliy ma’naviyatni takomillashtirish va yuksaltirishning asosiy yo‘llaridandir.

Zamonaviy axborot texnologiyalari tilning funktsional imkoniyatlaridan foydalanish borasida benihoya keng qulayliklar eshigini ochdi. Kompyuter tarjimasi, tahriri, tahlili, elektron lug‘atlar va tezaurus(til xazinasi)lar fikrimiz tasdig‘idir. Ayniqsa, zamonaviy elektron lug‘atlar tuzish va ulardan foydalanish madaniyatini shakllantirish til imkoniyatini kengaytirishda muhim ahamiyatga ega.

Respublikamizda har bir sohada axborot-texnologiyalaridan unumli foydalanishga, dasturiy ta’minot asosida tuziladigan loyihalarga katta e’tibor qaratilmoqda. Bu borada O‘zbekiston Respublikasi Birinchi Prezidentining bir qancha qarorlari va farmonlari, farmoyishlari qabul qilingan. Jumladan:

- O‘zbekiston Respublikasi Prezidentining 2002-yil 30-maydagi “Kompyuterlashtirishni yanada rivojlantirish va axborot-kommunikatsiya texnologiyalarini joriy etish to‘g‘risida”gi farmoni;
- O‘zbekiston Respublikasining 2003-yil 11-dekabrdagi “Axborotlashtirish to‘g‘risida”gi qonuni;
- O‘zbekiston Respublikasi Prezidentining 2012-yil 1-fevraldagi “Zamonaviy axborot-kommunikatsiya texnologiyalarini joriy etish va rivojlantirish chora-tadbirlari to‘g‘risida”gi PQ-1730 sonli qarori va shu kabilar misol bo‘la oladi.

Birinchi Prezidentimiz I.A.Karimov 2016-yil 13-maydagi PF-4749 sonli Farmoniga binoan Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetining tashkil etilishi o‘zida bir qator vazifalarni mujassamlashtirdi. Universitetning asosiy vazifalari sirasiga Farmonda quyidagilar ham kiritilgan:

- ✓ ona tilimizning internet jahon axborot tarmog‘ida munosib o‘rin egallashini ta’minlash, uning kompyuter uslubini, o‘zbek tili va dunyodagi yetakchi xorijiy tillar asosida tarjima dasturlari va lug‘atlar, elektron darsliklar yaratish bilan bog‘liq ilmiy-metodik ishlanmalar, amaliy tavsiyalar tayyorlash va bu borada erishilgan natijalarni amaliyotga keng tatbiq etish.

Bugungi kunda amaliy tilshunosligimiz oldida turgan eng muhim masalalardan biri – o‘zbek tilining milliy korpusini yaratishdir. Bunday mas’uliyatni, albatta, biz yoshlar o‘z zimmamizga olishimiz kerak. Zero, yosh avlodni tarbiyalashga qaratilgan davlat siyosati borasida Prezidentimiz Shavkat Mirziyoyev shunday deydi: “Yoshlarimizning mustaqil fikrlaydigan, yuksak intellektual va ma’naviy salohiyatga ega bo‘lib, dunyo miqyosida o‘z tengdoshlariga hech qaysi sohada bo‘sh kelmaydigan insonlar bo‘lib kamol topishi, baxtli bo‘lishi uchun davlatimiz va jamiyatimizning bor kuch va imkoniyatlarini safarbar etamiz”.[1]

Milliy korpus milliy til xazinasi demakdir. Undan lingvist, leksikograf, kompyuter lingvistlari, dasturchi, muharrir, tarjimon, jurnalist, noshirlar, olim, o‘qituvchi, ta`lim oluvchilar va boshqa har qanday soha mutaxassisi keng foydalanadi.

Til korpuslari — til bo‘yicha tadqiqot va amaliy topshiriqlar yechimi uchun inkor etib bo‘lmas ish quroli. U oddiy elektron kutubxonadan farqlanadi. Elektron kutubxonaning maqsadi — xalqning ijtimoiy-siyosiy, ma`naviy, iqtisodiy hayotini aks ettiruvchi badiiy va publitsistik asarlarni nisbatan to‘liq qamrab olish. Elektron

[1] Mirziyoyev Sh.M. Erkin va farovon, demokratik O‘zbekiston davlatini birgalikda barpo etamiz. -Toshkent: O‘zbekiston, 2016. –B. 14.

kutubxona matnlari til nuqtayi nazaridan ishlov berilmaganligi sababli tadqiqotlar uchun noqulaylik tug`diradi. Chunki elektron kutubxona ilmiy tadqiqot materiali bazasini tayyorlash maqsadida tuzilmaydi, balki milliy ma`naviy merosni jamlashni maqsad qilgan bo'ladi. Til korpusi esa elektron kutubxonadan farqli o'laroq, tilni o'rganish va tadqiq qilish uchun zarur, foydali va qiziqarli matnlarni to'plashni nazarda tutadi.

Mavzuning o'rganilish darajasi. Dunyo tilshunosligida korpus lingvistikasi yo'nalishi va til korpuslarini yaratish ishlari bo'yicha talay ishlar qilingan. Ko'pgina mamlakatlarda XX asrning 80-yillaridan boshlab bunday korpuslar tuzila boshlandi. Ular turli maqsad va vazifalarga xizmat qiladi. Buyuk Britaniyada Ingliz tili Banki (Bank of English) hamda Britaniya milliy korpusi (British National Corpus, BNC), Rossiyada Rus tilining milliy korpusi loyihalari ishlab chiqildi. Masalan, Rus tilining Milliy Korpusi hajmi hozirgi kunda 149 mln so'zdan iborat. Keyingi yillarda Internet tizimining rivojlanishi virtual matnlar korpusi yuzaga kelishiga olib keldi. Ya'ni Internetdagi qidiriv saytlari, elektron kutubxonalar, virtual ensiklopediyalar korpus vazifasini bajarmoqda. Korpusning janri va tematik rang-barangligi Internetdan foydalanuvchining qiziqishlariga bog'liq. O'zbek tilida milliy korpus yaratish ishlari hanuz orqaga surilmoqda. O'zbek tilida ko'plab lug'atlar yaratilgan, lekin til korpusi va uning nazariy asoslarini ishlab chiqish bo'yicha hali jiddiy tadqiqot qilingani yo'q. Buxorolik mustaqil tadqiqotchi Shahlo Xamrayeva bugungi kunda muallik korpusiga oid tadqiqotni amalga oshirmoqda. Uning tadqiqotining oxiriga yetishi bizga milliy korpus yaratish mezonlari uchun o'zbek tilida yaratilgan muhim manba sifatida munosib o'ringa ega bo'lardi.

Masalaning qo'yilishi. Bitiruv malakaviy ishida korpus lingvistikasining mohiyati, bu sohadagi xorijiy tajriba va bulardan kelib chiqqan holda o'zbek tilining milliy korpusini yaratish mezonlarini ishlab chiqish masalalari ko'rib o'tiladi.

Bitiruv malakaviy ishning maqsadi va vazifalari. Mazkur bitiruv malakaviy ishining maqsadi kelgusida yaratiladigan o'zbek tilining milliy korpusi

tamoyillarini belgilashdan iborat. Ushbu maqsadga erishish uchun quyidagi vazifalar bajarildi:

- mavzuga oid o‘zbek va xorijiy tillardagi adabiyotlar o‘rganib chiqildi;
- korpus lingvistikasi mohiyati o‘rganildi;
- dunyoda yaratilgan til korpuslarga oid ma’lumotlar to‘plandi;
- til korpusiga bo‘lgan talab va ehtiyoj aniqlandi;
- xorijda yaratilgan til korpuslarining xususiyatlari o‘rganildi;
- jahon tajribasiga tayangan hamda o‘zbek tilining tabiatini hisobga olgan holda o‘zbek tilining milliy korpusini yaratish tamoyillari ishlab chiqildi.

Ishning tuzilish tartibi. Tadqiqot kirish, ikki bob, har bir bob ichida uchtadan fasl, xulosa, foydalanilgan adabiyotlar va ilovani o‘z ichiga oladi.

I. KORPUS LINGVISTIKASI TAVSIFI

1.1. Til korpusi va uning mohiyati, korpusga bo‘lgan ehtiyoj

Korpus — til birliklarining xususiyatlarini aniqlash maqsadida qidiruv dasturiga bo‘ysundirilgan matnlar majmui, tabiiy tildagi elektron shaklda saqlanadigan yozma yoki og‘zaki, kompyuterlashtirilgan qidiruv tizimiga dasturiy ta’minot asosida joylashtirilgan on-line yoki off-line tizimda ishlaydigan matnlar jamlanmasi.

Korpus tilshunosligi esa kompyuter tilshunosligining tarkibiga kiruvchi tilshunoslik bo‘limidir, u korpus tuzilishining umumiy tamoyillarini ishlab chiqishda, matnlar korpusiga kompyuter texnologiyasini qo‘llash orqali loyihalar yaratishda ishtirok etadi. Tilshunoslik yoki til nuqtayi nazaridan matnlar korpusining tanasi katta, mashinadan o‘qiladigan shaklda ko‘rinadigan, yagona, tizimli, belgili, filologik jihatdan malakali til majmuasi hamda tilshunoslikka oid muayyan ma’lumotlar beruvchi baza sifatida tushuniladi. Hozirda “korpus” tushunchasini anglatishda ko‘plab ta’riflar mavjud. Masalan, E. Finegan: “Korpus - bu odatda, kompyuter o‘qiy oladigan formatda bo‘lgan va bizga matn ishlab chiqilgan vaziyat, informatsiya beruvchi, muallif, adresat yoki auditoriya haqidagi ma'lumotni o'z ichiga olgan matnlar to'plamidir”, - deydi. Turli umumiy ma’lumot beruvchi ijtimoiy saytlar korpusni statistik analiz hamda farazlar tekshiruvi asosida anglangan sohalarda uchrovchi til qoidalari va hodisalarini asoslay oladigan katta hajmdagi, tizimli matnlar to‘plami (endilikda odatiy elektron shaklda) sifatida ta’riflaydi.[2]

Til korpusi ma’lum tilning belgilangan davrdagi, xilma-xil janr, rang-barang uslub, hududiy hamda ijtimoiy variantdagi matnlarning elektron shaklli maxsus dasturiy ta’minot asosidagi yig‘indisidir. Korpus matnlar massividan iborat bo‘lib, bu matnlar oddiy elektron kutubxonadan farq qiladi. Korpusdagi matnlar maxsus qo‘shimcha ma’lumot bilan boyitilgan va lingvistik tadqiqot uchun asos vazifasini

[2] Захаров В.П. Корпусная лингвистика. – Иркутск, 2011.С. 7.

o‘taydi. Shunga asoslanib aytish mumkinki, til korpuslari, avvalo, tilshunos uchun kerak. Zero, korpussiz bugungi kun nazariy va amaliy filologiyasini tasavvur etish qiyin. Tilshunoslikka oid tadqiqotlarda dalil bilan ish ko‘riladigan hollarda o‘sha faktlar yig‘ilishi va sistemaga solinishi lozim. Bunday katta hajmli ishni bajarishda korpus tilshunos uchun vaqt va mehnatni tejaydigan bebaho ish qurolidir. Aslida korpus texnik jarayonni tezlashtiruvchi vosita bo‘libgina qolmay, ma’lum til zamonaviy shaklining axborot tizimi bo‘lib, kutilmagan savollarga ham javob bera oladigan, tilshunos oldiga avval qo‘ymagan dolzarb muammolarni qo‘ya oladigan tizimdir.

YUNESKO hisob-kitobiga ko‘ra, bugungi kunda mavjud 6 mingtacha tilning qariyb yarmi yaqin vaqtda o‘zining oxirgi sohiblaridan ham ajralishi mumkin. O‘lik tilga aylanayotgan bu tillarni saqlab qolish insoniyat oldidagi eng o‘tkir muammolardandir.

1999-yil 17-noyabrda YUNESKO Bosh anjumani e’lon qilgan Xalqaro ona tili kuni 2000-yilning fevralidan boshlab har yili lisoniy va madaniy taraqqiyot hamda ko‘p tillilikni ta`minlashga ko‘maklashish maqsadida nishonlab kelinadi. BMT Bosh assambleyasi o‘z rezolyutsiyasida 2008-yilni Xalqaro tillar yili deb e`lon qildi. [3]

Korpusni elektron kutubxonadan ajratib turuvchi birinchi omil undagi matnning xususiyati va qo‘shimcha ma’lumot bilan boyitilgani hisoblanadi hamda bu belgi korpusning alohida qismi — korpus birliklariga yozilgan izohni tashkil etadi. Foydalanuvchiga biror so‘z kerak bo‘lsa, buni odatiy matn muharriri ham topib beradi. Lekin matndagi til hodisasining ma’nosi, mazmuni va tuzilishini «tushunadigan» dasturiy tizim bilan ishlash juda afzal va qulay. Til birligini qidirish, kerak bo’lsa, bunday dasturiy ta’minot, ya’ni korpus tadqiqotchi yoki foydalanuvchiga juda katta yordam berishi mumkin. Tadqiqotchi o‘z ishi uchun misollar topish, ularni kartotekaga (kompyuter texnologiyalari rivojlanishidan oldingi davrda) ko‘chirishga oylab, ba’zan yillab vaqt sarflagan bo‘lsa, bugun dunyo til korpuslari yordamida sanoqli daqiqada yuzlab misollar topish, ular ustida

[3] Mengliyev B va boshqalar. O‘zbek tilining milliy korpusi//Ma’rifat, - Toshkent, 2018.

ishlash imkoniga ega bo‘ldi. Maxsus qidiruv tizimi korpusdan ma’lumot olishga mo‘ljallangan bir qancha dasturdan iborat. U statistik axborot va qidiruv natijasini foydalanuvchiga qulay shaklda taqdim eta oladi. Tilda qanday jarayon kechayotganini aniq tasavvur qilish uchun korpus qamrovini yanada kengaytirish, nafaqat yozma, balki og‘zaki nutq materialidan ham foydalanish maqsadga muvofiq. Bunday korpus yordamida taraqqiyot natijasida tilda sodir bo‘lgan va kutilayotgan o‘zgarish haqida aniq xulosa chiqarish mumkin.

Korpus bo‘yicha qidiruv foydalanuvchiga quyidagilarni aniqlash imkonini beradi:

1) Belgilangan til birligining turli qo‘llanishlardagi barcha shaklli ko‘rinishini;

2) Tilning lug‘at tarkibidagi o‘rni va variantlarini;

3) Belgilangan so‘z bilan birikish imkoniyatiga ega so‘zlar ro‘yxatini;

4) U yoki bu muallifning ayni so‘zdan foydalanish chastotasi yoki statistikasini;

5) So‘zning o‘z va ko‘chma ma’nolarini;

6) So‘z qo‘llanishining yashirin modeli(imkoniyati)ni;

7) Til taraqqiyotining turli davrida qo‘llanish holatini.

Korpuslar turli maqsadlarda, turli sohalar bo‘yicha tuzilishi mumkin. Dunyo tillari bo‘yicha yaratilgan korpuslarning tasnifi quyidagi jadvalda keltirildi:

Tasnif belgisi	***Korpus turlari***
Nutq turiga ko‘ra	Og‘zaki, yozma, aralash
Matn tiliga ko‘ra	Ruscha, inglizcha va hokazo
Parallelligiga ko‘ra	Bir tilli, ikki tilli, ko‘p tilli
Matnning ixtisoslashuviga ko‘ra	Badiiy, dialektal, suhbat, terminologik, aralash

Janriga koʻra	Adabiy, folklor, dramatik, publitsistik
Kirish usuliga koʻra	Erkin, pulli, yopiq
Maqsadiga koʻra	Tadqiqiy, tavsifiy
Izohlanishiga koʻra	Qoʻshimcha izohli, izohsiz
Oʻzgaruvchanligiga koʻra	Oʻzgaruvchan, turgʻun
Strukturasiga koʻra	Markaziy, arxiv, mahalliy
Izoh xususiyatiga koʻra	Morfologik, sintaktik, semanti
Matn hajmiga koʻra	Toʻliq matnli, lavha matnli
Xronologik jihatiga koʻra	Sinxron, diaxron
Umumiyligiga koʻra	Umumiy, yakka muallifli, sohaviy

Mavjud korpuslar tarkibidagi matnlarning nisbatiga qarasak, badiiy adabiyot hissasi 40 foizni tashkil etganiga guvoh boʻlamiz. Uning tarkibiga memuar asarlar ham kirib ketadiki, bu janr til xususiyati badiiy va publitsistik uslub oralig‘ida boʻlib, jonli tilni oʻrganish uchun ancha qulay. Yevropa tillari korpuslarida badiiy adabiyot materiali 20 foizni tashkil etadi. Masalan, zamonaviy yozuvchilar til xususiyatini oʻrganishga bag‘ishlangan 20 dan ortiq tadqiqot mavjud boʻlsa-da, ular hali toʻlaligicha bu muammoni oʻrganib boʻldi, deyish qiyin. Chunki alohida yozuvchi asarining til xususiyatidagi oʻzgarishga hali tildagi yangi hodisa deb qarab boʻlmaydi. Ana shunday jarayonlarni kuzatish va tadqiq etishning eng qulay vositasi — til korpusi.

Korpusning boshqa sohalardagi ahamiyati. “Korpus tor mutaxassislar uchungina keraklimi yoki yana boshqa soha vakillari hamundan foydalansa boʻladimi?” degan savolga korpus lingvistikasiga oid qator adabiyotlarda korpusning ahamiyati toʻg‘risida soʻz borar ekan, uning qoʻllanilish doirasining

ancha kengligi ta’kidlanadi. Korpus foydalanuvchilari doirasi biz o‘ylaganimizdan ko‘ra ancha keng bo‘lib, quyida shularni sanab o‘tamiz.

Lug‘at va grammatikadan faqat tilshunos foydalanmaganidek, korpus gumanitar fanlar tadqiqotchisi, adabiyotshunos hamda tarixchi uchun ham birdek zaruriy baza, deyish mumkin.

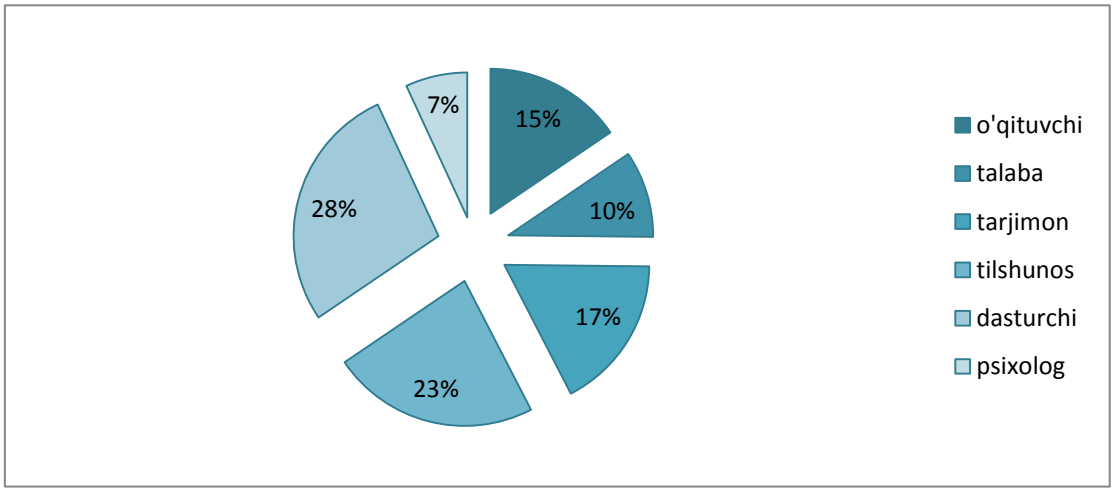
O‘qituvchi uchun korpus — tengsiz xazina. Masalan, u topshiriqni tayyorlash uchun faqat badiiy adabiyotdan (yoki badiiy bo‘lmagan matndan: sport, avtobiografiya yoki shunchaki ma’lum ijodkor asari bilan cheklanmoqchi) foydalanmoqchi. Unda «Mening korpusim» ilovasi orqali o‘ziga kerakli matnni ajratib olish imkoni mavjud, ya’ni misollar sirasini chegaralaydi. Korpus turli janr va uslubdagi matnlarni qamrab olgani sababli har qanday talabni qondirish imkoniga ega. Misol uchun, jurnalistika fakulteti talabalari uchun yozilgan darslikda publitsistik matnlar ancha eskirgan. Har kuni muntazam boyitiladigan korpusdan yangi misollar olish, ularni talabaga taqdim etish yoki unga ham shunday topshiriq berish, albatta, ta`limni hayotga yaqinlashtiradi. Ko‘pincha qayta-qayta nashr etilgan darslikda misollar eskiligicha qolib ketadi. Bugungi kun uchun mavzu eskiradi, talabalar bilimni davrga hamohang o‘zlashtirmaydi.

Bugun filolog yoki jurnalist bo‘lib yetishayotgan mutaxassisga tilimizning barcha nozik qirralarini his etishi uchun faqat eski — 20-30 yil oldingi matn namunasi yetarli emas, u bugun yozilgan barcha soha, uslub va janrdagi asarlar bilan ishlashi lozim. Shundagina ta’lim ijtimoiy buyurtmani bajargan hisoblanadi.
Til korpusidan o‘qituvchi, talaba va maktab o‘quvchisi ham ancha unumli foydalanishi mumkin. Chunki faqat korpus orqali juda osonlik bilan kam ishlatiladigan so‘z, ibora va birikmani topish, qo‘llanishi va yozilishi(orfografiyasi)ni o‘rganish mumkin. Ta’kidlash joizki, til korpusida til grammatika va darslik muallifi tavsiflaganidek emas, balki jamiyatda qanday yashasa va ishlasa, shunday aks etadi. Bu esa o‘quvchining o‘zini o‘rab turgan muhit — umumxalq tili va adabiy tilni o‘rganishida eng sermahsul vosita bo‘lib xizmat qiladi.

Til korpusiga eng ko‘p ehtiyoj sezuvchi mutaxassis bu — matnni avtomatik qayta ishlash (masalan, tarjima dasturi) va turli qidiruv tizimlari bilan ishlaydigan dasturchi. Chunki u tabiiy til bilan ish ko‘radi hamda ushbu tilda yozilgan barcha matnlar strukturasini (tabiiy va jonli tilda, ilmiy grammatika va darsliklardagi misollarga tayanib emas!) mukammalroq «tushunishi», his etishi lozim.

Til korpusiga ehtiyoj sezadigan mutaxassislardan yana biri — kundalik ish jarayonida yozma va og‘zaki nutq jozibasiga tez-tez murojaat etuvchilar: gazeta va jurnal muharriri, jurnalist, radio va televideniya xodimlari. Chunki bu mutaxassislar muayyan so‘z, ibora yoki gapning qo‘llanish holati, darajasi, kim, qachon ilk marta bu gapni qo‘llagani, qanday uslub uchun xoslanishini bilishga grammatika bilan shug‘ullanuvchi olimdan ko‘ra ko‘proq ehtiyoj sezadi. Korpusdan tashqari hech qanday axborot banki bu kabi savollarga zudlik bilan javob berishi mumkin emas. Shuning uchun korpus lingvistikasi tadqiqotchilari korpuslar jurnalist, muxbir, muharrir, o‘qituvchi hamda dasturchi uchun maxsus yaratila boshlangan, degan xulosaga ham kelishadi. Biz bugungacha darslik, ilmiy grammatika va lug‘atga qanday tayangan bo‘lsak, zamonaviy soha vakillari korpusga shu qadar ehtiyoj sezishadi, undan foydalanishadi.

BNC (Britaniya milliy korpusi) foydalanuvchilarining bir kunlik monitoringi



Xulosa sifatida suni aytish mumkinki, korpusning ijtimoiy ahamiyati juda keng qamrovli. Korpusdan lingvistik va etno-psixolingvistik tadqiqotlarda, ona tili, adabiyot, xorijiy til ta’limida, matnga ishlov berish va tarjima dasturlarini tuzishda foydalanish mumkin. Korpus tilshunos, tarjimon, o‘qituvchi, dasturchi, gazeta va jurnal muharriri, jurnalist, radio va televideniya xodimi, umuman kundalik faoliyat jarayonida “ish quroli so‘z bo‘lgan” har qanday kishi uchun zamonaviy axborot vositasining bir ko‘rinishidir.

1.2. Korpus lingvistikasining tilshunoslik bo‘limlariga munosabati

Korpusning keng ko‘lamli bo‘lishi ma’lumotning o‘ziga xosligini kafolatlaydi. Til hodisalarining barcha qirrasini to‘liq namoyish etishni ta’minlaydi. Turli tipdagi ma’lumot til korpusida qo‘llanish shakliga ko‘ra joylashadi. Bu esa uni har tomonlama va xolis o‘rganish uchun asos bo‘ladi. Bir marta tuzilgan va tayyorlangan axborotlar massivi bir necha tadqiqotchi tomonidan ko‘p marta, turli maqsadda ishlatilishi mumkin. Korpus — tilni tadqiq etish (til birligining o‘zgarishi, eskirishi, yangilarining paydo bo‘lishi, ma’nosining kengayishi va torayishi; yangi iboralarning paydo bo‘lishini kuzatish), o‘rganish, an’anaviy va zamonaviy lug`atlar tuzishda keng imkoniyatli dasturlashtirilgan tizim. Korpussiz bugungi kun nazariy va amaliy filologiyasini tasavvur etib bo‘lmaydi.

Tilshunoslikka oid tadqiqotda fakt bilan ish ko‘riladigan hollarda material yig‘ilishi, sistemaga solinishi lozim. Bunday katta hajmli ishni bajarishda korpus vaqt va mehnatni tejaydigan ish quroli vazifasini bajaradi. U texnik jarayonni tezlashtiruvchi vositagina emas, muayyan til zamonaviy shaklining axborot tizimi bo‘lib, kutilmagan savolga ham javob bera oladigan, til hodisasi bilan shug‘ullanadigan, soha oldiga avval qo‘yilmagan dolzarb muammolarni qo‘ya oladigan tizim.

Korpus katta hajmli lug‘atlarni tuzish uchun manba vazifasini o‘taydi. Vaqt o‘tishi bilan korpus turli lingvistik yo`nalishlar uchun ahamiyatli bo‘lishi bilan

kuchli informatsion resursga aylandi. Korpus asosida kompyuter yordamida lug‘atlar avvalgiga nisbatan tez tuziladi va qayta ishlanadi. Shu yo‘l bilan ish boshlanishidan tugash jarayonigacha (nashrgacha) tilni aks ettirib turadi. Lug‘at maqolasi «eskirish»ga ulgurmaydi. Tilning zamonaviy holatini aks ettiruvchi korpus turli davrlarda yashab ijod etgan ijodkorlarning mualliflik korpuslari uchun asos bo‘lib xizmat qiladi.

So‘zning qo‘llanish davri va chastotasini aniqlashda hech qanday vosita korpusga tenglasha olmaydi. Korpus lingvistikasining keyingi taraqqiyot bosqichida statistik tadqiq metodi kompyuter tarjimasi, nutqni sintezlash va tanish, aniqlash, orfografik tekshiruv kabi lingvistik amallarni bajarishda qo‘l kela boshladi. Rus va Yevropa tilshunosligida til bo‘yicha o‘tkaziladigan barcha tadqiqotlarning korpusga asoslanishi odatiy holga aylangan va hatto ba’zi tadqiqotlarda majburiy ham sanaladi. [4]

Gap tuzilishini o‘rganishdagi ahamiyati tilning jonli qurilishini, so‘zlarning o`zaro birikish imkoniyatini tahlil qilishda ko‘rinadi. An’anaviy tilshunoslikdagi misolni olishda badiiy asar tiliga tayanish tajribasidan ko‘ra korpusga tayanish misolning bugungi kun uchun ham ishonarliligini ta’minlaydi. Korpus tarjima dasturlarini yaratishda katta ahamiyatga ega.

Korpus asosida ish ko‘radigan eng birinchi soha leksikografiya bo‘lib, katta hajmli lug‘atlarni tuzish uchun asosiy va takrorlanmas manba sanaladi. Vaqt o‘tishi bilan korpuslar turli lingvistik yo‘nalishlar uchun ahamiyatli bo‘lishi bilan kuchli information resursga ayandi. Chunki korpus leksikografiya sohasi uchun boy manba hisoblanib, ular asosida kompyuter yordamida lug‘atlar avvalgiga nisbatan tezlik bilan tuziladi va qayta ishlanadi. Shu yo‘l bilan ish boshlanish va tugash jarayonigacha (nashrgacha) tilni aks ettirib turadi, eskirishga ulgurmaydi.

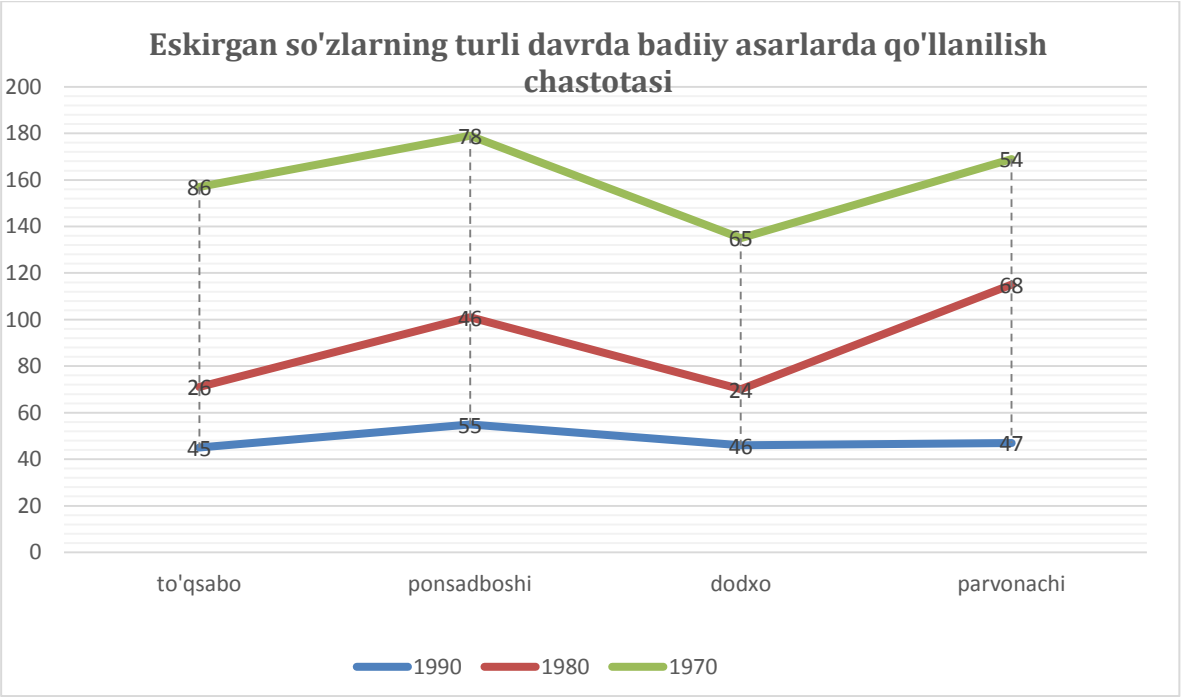
Kompyuter leksikografiyasini elektron matnlar korpusi yoki parallel matnlar korpuslarisiz tasavvur qilish mumkin emas. Chunki dunyodagi barcha zamonaviy, so‘nggi lug‘atlar korpusga asoslangan bo‘lib, ular misollarining haqiqiy, ishonarliligi bilan baholanadi. Chunki korpusda til jamiyatda qanday yashasa,

[4] Xamrayeva Sh. Til korpusining leksikografik ahamiyati // O‘zMU, - Toshkent, 2017.

shunday aks etadi, natijada lug‘atdagi misol ishonarli hamda asosli bo‘ladi. **Matnlar korpusi («corpus» lotincha «tana» degan ma'noni anglatadi)** - bu elektron holda saqlanadigan ma'lum til birliklari bo'lib, ular tilshunoslar uchun turli xil muammolarni hal etish uchun tatbiq etishda va turli yo‘nalishdagi tadqiqotlar uchun zaruriyatga qarab turli shakllarda tuziladi. Bular fonema, grafema, morfemalardan tortib undan kattaroq birliklar: leksema, gap va matnlardan (badiiy yoki ilmiy asar, gazeta va jurnal matnlari) tashkil topishi mumkin. Ularning qay tarzda saqlanishiga qarab maxsus dasturlar yordamida har bir kerakli so‘z yoki so‘z birikmasi uchun darhol uning qo‘llanishi bo‘yicha misollar topilishi, imlo bo‘yicha variantlari, sinonimik qatorlari topilishi mumkin. Matnlar korpusiga oid ilmiy tadqiqotlar salmog‘ining ko‘payishi natijasida tilshunoslikda **korpus lingvistikasi** yo'nalishi shakllandi.[5]

Korpusning leksikologiya sohasidagi ahamiyati shundan iboratki, so‘zning qo‘llanish davri va chastotasini aniqlashda hech qanday vosita korpusga tenglasha olmaydi. Korpus asosida ma’lum so‘zning chastotasini aniqlash uchun berilgan qidiruv natijasida diagramma va grafiklar yordamida so‘zning tartib raqami uning chastotasiga teskari proparsional bo‘ladi, chunki ikkinchi tartib raqamida joylashgan so‘z birinchi raqamli so‘zga nisbatan kamroq, to‘tinchisi uchinchisiga nisbatan kamroq ishlatilishi aniq. Birorta chastota lug‘ati korpuscha aniq ma’lumot berolmaydi, chunki til doim o‘zgarishda bo‘lib, so‘zning chastotasi ham nisbiydir. Korpusning Sipfa qonuniyati deb ataladigan bunday amaliyoti asosida - “Til – eng kamyob hodisalar yig‘indisi”, degan xulosaga kelishga asos bo‘ladi. Til korpusi yordamida o‘tkazilgan ilk lingvistik tadqiqotlar turli birlikning tilda qo‘llanish chastotasini statistik aniqlashni maqsad qilardi. Korpusga ma’lum so‘zning turli davrda asarlarda qo‘llanilish chastotasini aniqlash bo‘yicha berilgan so‘rov natijasi taxminan quyidagi shaklda namoyon bo‘ladi. Natijani namoyish etish shaklini foydalanuvchi istagiga ko‘ra turli shaklada tanlashi mumkin:

[5] Rahimov A. Kompyuter lingvistikasi asoslari. – T., 2011



Bundan tashqari, korpus tilning lug‘at boyligida bo‘layotgan o‘zgarish (neologizmning paydo bo‘lishi va yo‘qolishi hodisasi)ni kuzatishning eng qulay vositasi sanaladi. So‘zlarning o‘zaro leksik-semantik birikish imkoniyati tahlilida korpus metodini qo‘llash yangi avlod lug‘at va grammatikalari, xususan turg‘un birikmalar lug‘atini yaratish imkonini berdi. Korpus yaratilishi va taraqqiy etishi, til korpuslari asri kirib kelishi bilan lug‘atshunoslarda so‘z qo‘llanish holatlariga oid juda katta matnlar to‘plami bilan ishlash imkoniyati paydo bo‘ldi. Korpus vositasida so‘zning serqirraligi, bir paytning o‘zida bir necha smantik kayorliqoriyalarga mansub bo‘la olishi, bu semantik farqni ajratib olish mumkinligi hqida Y.V. Nedovshina: “Leksikografiya bo‘yicha tadqiqotlar semantika borasidagi izlanishlar bilan chambarchas bog‘liq. Korpusda u yoki bu lingvistik birlikning qurshovini kuzata turib, ushbu birlikni xarakterlovchi ma’lum semantik belgilarni aniqlash mumkin. So‘z bir paytning o‘zida bir necha semantik kayorliqoriyaga mansub bo‘lishi mumkin, shuning uchun so‘zning u yoki bu kayorliqoriyaga yorliqishlilik darajasiga qarab fikr yuritish lozim. Daraja esa turli kayorliqoriya bo‘yicha taqsimlanish chastotasini anab o‘tish yo‘li bilan

aniqlanadi”[6] – deb yozadi. Demak, korpus yordamida so‘zning semantik kayorliqoriyalarga mansubligi, har bir kayorliqoriyadagi ma’nosi bilan tanishish mumkin. Til korpusi yordamida o‘tkazilgan ilk lingvistik tadqiqotlar turli birlikning tilda qo‘llanish chastotasini statistik aniqlashni maqsad qilardi. Ko‘pincha, bunday element so‘z, ba’zida grafema, morfema, so‘z birikmasi ham bo‘lardi. Korpus lingvistikasining keying taraqqiyot bosqichida statistik tadqiq metodi kompyuter tarjimasi, nutqni sintezlash va tanish, aniqlash, orfografik tekshiruv kabi lingvistik amallarni bajarishda qo‘l kela boshladi. Korpusdan foydalanish leksik birlikni kontekstda o‘rganish imkonini beribgina qolmay, so‘zshakl, lemma, grammatik kayorliqoriyalar chastotasi, ularning birikish xususiyati haqida ma’lumot olish uchun ham muhim. Ibora yoki turg‘un birikmalar semantik jihatdan bo‘linmas birlikni tashkil etgani uchun leksikografiyada matnga avtomatik ishlov berish jarayonida buni inobatga olish juda muhim. Korpus materiali asosida statistik metod orqali qaysi so‘zlar doim birgalikda qo‘llanishi, shuning natijasida turg‘un birikmaga qanchalik aloqador ekanligini aniqlash mumkin.[7]

Uslubiyatni o‘rganishda matnning o‘ziga xosligini tahlil etishda ham korpusga tayanish mumkin. Bu amaliyotni bajarishda matnlarning statistik holati tahlili (matndagi jumlaning uzunligi, bir so‘zning boshqa so‘z bilan birikish holatining doimiy yoki noodatiyligi) aniqlanadi. Bunday usulda korpus yordamida yozma nutq bilan birga og‘zaki nutqni ham o‘rganish mumkin. Masalan, rasmiy ish qog‘ozlarida hujjat matnida shartnoma bo‘yicha talablar shakli ma’qulmi yoki shartnoma talablari konstruktsiyasi to‘g‘rimi? Korpus bu savolga javoban sanoqli daqiqalarda qaysi shakl ko‘proq qo‘llanishini aniqlab beradi. Shunday qidiruv natijasida rus tili milliy korpusida shartnoma bo‘yicha talablar konstruksiyasi 40 marta, shartnoma talablari shakli 3 marta qo‘llangani to‘g ‘risida statistik ma’lumot

[6] Недовшина Е.В. Программа для работы с корпусами текстов: обзор основных корпусных менеджеров. Работа с системой DDC // Языковая инженерия: в поиске смыслов. (электронный ресурс).
[7] Захаров В.П. Корпусная лингвистика. Учебно-методическое пособие. – СПб: БИ, 2005. – С. 48.

beriladi. Bu konstruksiyaning qo‘llanishi, albatta, rus tiliga xos. Ehtimol, u o‘zbek tilida boshqacha natija berar. Korpusimiz bo‘lmagani tufayli bu savolga aniq javob bera olmaymiz.

Korpus juda muhim bo‘lgan va faol qo‘llanadigan soha — lingvodidaktika. U ona tili va xorijiy tilni o`rganishda birdek ahamiyatli. Tilni o`rgatishda lug`at boyligini ko‘rsata olish, so‘zning qo‘llanish imkoniyatini u yoki bu grammatik qurilish orqali tushuntirish uchun misollar keltirishda korpus juda zarur. Til ta`limi uchun muhim bo‘lgan misolning doimiy yangilanishi, buni aks ettirish xususiyati hamda imkoniyati faqat korpusga xos. O‘qituvchi yangi, ishonarli, cheksiz hamda xilma-xil misollarni shu erdan topa oladi. Topshiriq va mashqlarni belgilashda qiynalmaydi. Bir necha daqiqada mavzu bo‘yicha yangi misollardan iborat mashq tayyorlay oladi. Korpusda mavjud matnlarni saralash, misolni barcha matnlardan emas, balki izlanuvchi uchun qiziqarli va kerakli bo‘lganidan ajratish mumkin. Hatto korpus matnlarini belgilangan davr (hatto aniq yiligacha), matnning aniq bir turi (masalan, reklama matni, ish hujjati yoki shunchaki bir necha muallif asarlari)ni tanlab olish imkoniyati mavjud. Turli mavzu va janrdagi matnlar bilan doimiy boyitilishi — korpusning asosiy xususiyatlaridan bo‘lib, ta’limda yo‘naltirib o`qitish imkoniyatini ochadi. Akademik litsey va kollejda ona tili barcha yo`nalish va mutaxassisliklarda o`qitiladi. Yo‘nalishlar mavzusi bir xil bo‘lishi mumkin, lekin har bir mutaxassislikda mavzuni tushuntirish uchun mos matnni tanlashda korpusda mavjud imkoniyat birorta manbada yo‘q. Aslida, akademik litsey va kasb-hunar kollejlari davlat ta’lim standartlarida maqsad egallanayotgan mutaxassislik bo‘yicha yozma va og‘zaki nutqni rivojlantirish ekani belgilangan. Ayni bir mavzu (masalan, so‘zning shakl va ma’no munosabatiga ko‘ra turi, atamashunoslik, toponim, qo‘llanish doirasi chegaralangan leksika va hokazo) turli yo‘nalishda alohida, ixtisoslashtirilgan matnlar asosida o‘rganilishi lozim. Shundagina til ta’limi o‘z maqsadiga erishadi.Til ta’limiga oid darslik va qo‘llanmalarda iqtisod, moliya, siyosat, huquqshunoslikka xos matnlarga duch kelmaymiz, misolni badiiy asardan olishga odatlanganmiz. Vaholanki, adabiy til badiiy asar tili degani emas. Til ta’limi

hayotdan uzilib qolgan. Kundalik turmushda oʻquvchi tuzadigan, ishlatadigan yoki tahlil etadigan matnlardan uzilib, mashq va topshiriqni jonli tildan olmaslikka koʻnikib qolganmiz.

N.R.Dobrushina rus tili bo`yicha darslik va qoʻllanmalarni tahlil qilarkan, shunday yozadi: «Tipik rus tili darsligida mashqdagi misollarning 60 foizi XIX asr, 20 foizi XX asr badiiy adabiyotiga oid; 10 foizi muallif tomonidan tuzilgan; zamonaviy OAV materialidan esa bitta ham misol yo`q. Albatta, darslikda mumtoz nasr namunasi berilishi tabiiy, ijobiy holat. Bugungi kun o`quvchisi o`zi uchun ancha notanish bo'lgan bu matnni o`qishni unutmasligi uchun ham kerak. Agar ta`limning biz o`qitayotgan shaxsga aniq qaratilishi va ta`lim maqsadiga erishishini, ta`lim va hayot o`rtasida favqulodda katta tafovut bo`lmasligini istasak, zamonaviy til(hozirgi adabiy til)ga murojaat etishimiz, darslikka faqat badiiy adabiyot emas, balki kundalik turmush tarzi va hayot bilan bog`liq matn namunalarini ham kiritishimiz kerak».

Korpus lingvistikasi mutaxassislari tajribasiga nazar solsak, korpus yordamida qanday amallar bajarilishi mumkinligiga guvoh boʻlamiz. Rus tilida *реагировать* fe'lining (несовершенный вид) совершенный koʻrinishining *прореагировать, отреагировать, среагировать* kabi shakllari mavjud**.** Ushbu old qoʻshimchalardan qaysi biri koʻproq qoʻllanishi, qaysi uslubdagi kontekstga xoslanishi, qaysi ravishlar bilan bogʻlana olishi, hozirgi rus tilida qanday tartibda (bir paytda yo ketma-ket) ishlatilishi, til taraqqiyotining turli davrida qoʻllanish chastotasining farqi kabi masalalarga korpus bir necha daqiqaga oydinlik kiritib beradi.[8]

Rus tilida ot soʻz turkumi мужской род birlik paradigmasida qoʻshimcha kelishik kayorliqoriyasi (второй родительный падеж) mavjud. *Сахар* soʻzi bu kelishikda *сахара* shaklida, ikkinchi shaklda esa *сахару* shaklida qoʻllanadi: *положите себе еще сахару*. Bunday turlanish shakli rus tiliga XVI-XVII asrda kirib kelgan, XVII-XVIII asrlarda eng koʻp qoʻllanish davri boʻlib, XIX asrda

[8] Плунгян В. Зачем мы делаем корпусы? // www.ruscorpora.ru

sekin nutqdan chiqa boshlagan. Hozirgi rus tilida ushb shakl yuzdan bir holatni tashkil etadi va kelishikning birinchi shaklida qo‘llanaveradi: *положите себе еще сахара*. Hozirgi rus tilida ushbu kelishik shaklining qo‘llanish chastotasi, til taraqqiyotining turli davrida, turli muallif nutqida ikkinchi родительный падеж ning qo‘llanish holatlari (*поднять с пола* èки *поднять с полу*) masalasi sanoqli daqiqalarda statistik va faktik asosda hal etiladi.[9] Bunday misollar sirasini ko‘paytirish, muammoni ko‘proq qo‘yish mumkin. Korpusning tilshunos oldidagi amaliy ahamiyatini aniqroq ko‘rsatish uchun shunday oddiy misol keltirildi. Rus va Yevropa tilshunosligida til bo‘yicha qilinadigan barcha tadqiqotlarning til korpusiga asoslanishi odatiy holga aylangan va hatto ba’zi tadqiqotlarda majburiy ham sanaladi. Rus tilshunoslari bu sohada AQSh, Yevropa, Yaponiyadan orqada qolganini afsus bilan tan olishadi, vaholanki, rus korpus lingvistikasi ancha taraqqiyot yo‘lini bosib o‘tib, bu borada rivojlangan mamlakatlarga tenglashyapti. Ularning fikricha, yaqin orada til o‘rganayotgan o‘quvchi yoki uning biror jihatini tadqiq etayotgan kishi bugungi kunda korpusdan foydalanishi aniqligini ko‘rsatadi.

Korpusning sintaksisni o‘rganishdagi ahamiyati tilning jonli qurilishini, so‘zlarning o‘zaro birikish imkoniyatini tahlil qila olishidadir. An’anaviy tilshunoslikdagi misolni olishda badiiy asar tiliga tayanish tajribasidan ko`ra korpusga tayanish misolning bugungi kun uchun ham ishonarliligini ta’minlaydi.

[9] Плунгян В. Зачем мы делаем корпусы? // www.ruscorpora.ru

I. MILLIY KORPUSLARNING YARATILISHI VA O‘ZBEK MILLIY KORPUSI

1.1. Milliy til korpusining paydo bo‘lishi va taraqqiyoti

Ochig‘i, korpusni insoniyat hali korpus lingvistikasi fani paydo bo‘lmasdan oldin, ya’ni XVIII asrdan tadqiq eta boshlashgan. Misol qilib oladigan bo‘lsak, Bibliyani tadqiq etish, lug‘atlar yaratish (Johnson, Oxford English Dictionary, Webster Dictionary), tillarni o‘qitish (chastotali lug‘at Thorndike'a, 1921), deskriptiv grammatika (Fries, 1940, Quirk, 1968) va boshqalar. Kvirk korpusi bir million so‘z birikmalarini o‘zida jamlagan bo‘lib, har biri o‘n yetti qator matndan iborat million kartotekadan iborat. Bu korpus oxirgi elektron shaklda bo‘lmagan korpus edi. Mazkur korpusning yaratilishiga 25 yil vaqt sarflangan. Kvirk korpusi 1989-yilda oxirgi ishlarini yakunlagan. Bu paytda texnologiyalar yuqori sur’atda rivojlanib ketayotgan edi. Buning natijasi o‘laroq korpus tezlikda elektron shaklga o‘tkazildi. Hozirda bu korpus Londondagi University kollejida saqlanmoqda.

Kompyuter yordamida yaratilgan korpuslarning asosiy davrlari:

1. 1960-yil, Braun korpusi (AQSh), bir million so‘zni qamrab oladi;
2. 1970-yil, LOB korpusi (Buyuk Britaniya, Norvegiya), bu korpus ham bir million so‘zni qamrab oladi;
3. 1980-yil, Rus tilining mashinalashgan fondi (Машинный Фонд русского языка);
4. Rus tilining Uppsal korpusi (Shvetsiya), bir million so‘zni qamrab oladi;
5. 1990-yil, Britaniya milliy korpusi (British National Corpus), milliy korpuslar (venger, italyan, xorvat, chex, yapon millatlari tillari) 100 million so‘zni o‘z ichiga oladi;
6. Ingliz tili majmui (The Bank of English), Birmingham (Collins Cobuild), 600 million so‘zdan iborat;
7. 2000-yil, Amerika milliy korpusi (American National Corpus), 100 million so‘zni o‘z ichiga oladi;

8. Amerikacha ingliz tilining zamonaviy korpusi (Corpus of Contemporary American English), 400 million so‘zdan iborat;
9. Rus tilining milliy korpusi (Национальный корпус русского языка), 140 million so‘zni o‘z ichiga oladi;
10. Gigaword corpora: ingliz, arab, xitoy tillarini qamrab olgan bo‘lib, 2 milliard so‘zdan tashkil topgan;
11. Oxford English corpus, mazkur korpus ham 2 milliard so‘zdan tashkil topgan;

Bugungi kunda har bir rivojlangan til o‘z korpusini yaratish ustida ish olib bormoqda. Eng mashhur va ko‘zga ko‘ringan korpuslarni yuqorida sanab o‘tdik. Kompyuter texnologiyasining tez sur’atdagi rivoji bundanda katta hajmdagi korpuslarni ilm ahliga taqdim etadi.

V.V.Rikovning yozishicha, korpus lingvistikasi atamasi munozarali masala. Chunki korpus lingvistikasi matnlar massivini yaratish masalasimi yoki korpus ma’lumotlari asosidagi lingvistikami? Amaliyotda korpus lingvistikasi deganda:

- birinchidan, korpus makur fan uchun nutq materiali;
- ikkinchidan, faoliyat natijasidir.

Xulosa qilib aytganda, quyidagilarga imkoniyat yaratdi:

1. Tilga oid ilgari qilingan tadqiqotlar ko‘lami va natijalarini kompleks holda aniqlash va tahlil qilish;
2. Yangi va keng ko‘lamda lingvistik tadqiqotlar olib boorish.

Korpus lingvistikasi fani o‘z obyekti va ish materialini o‘zi yaratdi, bu esa uni mustaqil lingvistik fan sifatida tan olishga asos bo‘ladi. Biz tadqiq etayotgan fanning asosiy maqsadi – til tizimini lingvistik tasvirlashdir.

Korpus lingvistikasining asosiy yo‘nalishlari:

Zamonaviy korpus lingvistikasining asosiy yo‘nalishlari quyidagilar:

- birinchidan, bu lug‘atlar yaratish hamda leksikografik tadqiqotlar olib borishdir, zamonaviy ingliz tilining barcha lug‘atlari korpusga asoslangan (Collins, Webster, MacMillan va boshqalar);

- ikkinchidan, korpuslarni o‘rganish orqali tillarning leksik tarkibi haqida aniq ma’lumotlar olish, so‘zlarning qo‘llanish chastotalarini tuzish. Korpusning leksikologiya sohasidagi ahamiyati shundan iboratki, so‘zning qo‘llanish davri va chastotasini aniqlashda hech qanday vosita korpusga tenglasha olmaydi. Korpus asosida ma’lum so‘zning chastotasini aniqlash uchun berilgan qidiruv natijasida diagramma va grafiklar yordamida so‘zning tartib raqami uning chastotasiga teskari proparsional bo‘ladi, chunki ikkinchi tartib raqamida joylashgan so‘z birinchi raqamli so‘zga nisbatan kamroq, to’tinchisi uchinchisiga nisbatan kamroq ishlatilishi aniq. Birorta chastota lug‘ati korpuscha aniq ma’lumot berolmaydi, chunki til doim o‘zgarishda bo‘lib, so‘zning chastotasi ham nisbiydir. Korpusning Sipfa qonuniyati deb ataladigan bunday amaliyotiga asosan, har bir tilda tez-tez ishlatiladigan so‘zlarni aniqlashning imkoniyati endi yuqori. [10]

Kompyuterda yaratilgan birinchi matnlar korpusi Braun korpusi (БК, inglizcha Brown Corpus, BC) hisoblanadi, u 1961-yilda Braun universitetida yaratilgan, har biri 2000 so'zli 500 ta matn fragmentini o'z ichiga oladi. 1970-yillarda 1 mln so'zni o'z ichiga olgan matnlar korpusi asosida rus tilining chastotali lug'ati yaratildi. 1980-yillarda Shvetsiyaning Upsala universitetida ham rus tilida matnlar korpusi yaratildi. Keyinchalik kompyuter leksikografiyasining rivojlanishi natijasida katta hajmli matnlar korpusiga ehtiyoj tug'ildi. Ya'ni 1 mln ta so'z elektron lug'atlar bazasi uchun yetarli emas. Shu asosda yirik hajmli matnlar korpusi yaratila boshlandi. Ko'pgina mamlakatlarda XX asrning 80-yillaridan boshlab bunday korpuslar tuzila boshlandi. Ular turli maqsad va vazifalarga xizmat qiladi. Buyuk Britaniyada Ingliz tili Banki (Bank of English) hamda Britaniya Milliy Korpusi (British National Corpus, BNC), Rossiyada Rus tilining Milliy Korpusi loyihalari ishlab chiqildi. Masalan, Rus tilining Milliy Korpusi hajmi hozirgi kunda 149 mln so'zdan iborat. Keyingi yillarda Internet tizimining

[10] Рыков В.В. Корпус текстов как новый тип словесного единства // Труды Междунар. семинара «Диалог-2003». М.: Наука, 2003. С. 22–23.

rivojlanishi virtual matnlar korpusi yuzaga kelishiga olib keldi. Ya'ni Internetdagi qidiriv saytlari, elektron kutubxonalar, virtual ensiklopediyalar korpus vazifasini bajarmoqda. Korpusning janri va tematik rang-barangligi Internetdan foydalanuvchining qiziqishlariga bog‘liq. Masalan, ilm-fan doirasida Wikipedia katta hajmdagi matnlar korpusi sifatida foydalanilmoqda.[11]

Ayniqsa, ona tili va chet tillarini o‘qitish va o‘rganish borasida korpusning ahamiyati beqiyos. Bugungi kunda dunyo miqyosida til o‘rgatish tizimi korpuslarga yo‘naltirilayotganligi ham – fikrimizning dalili. Shuning uchun ta`lim korpuslari, sheva matnlari korpuslari, poetik matnlar korpusi, og`zaki, ilmiy, rasmiy matnlar korpusi, parallel korpus kabi qator mikrokorpuslarning tuzilayotganligi ahamiyatli. Ingliz, nemis, fransuz, rus tillarini xorijiy til sifatida o‘qitish masalasi metodikada alohida tadqiq etilmoqda. Aynan til o‘rgatishni maqsad qiluvchi korpuslar ham mavjud bo‘lib, «Учебный корпус русского языка», «Learner corpus of English» shular jumlasidan. Xorijiy til vakillari bilan ishlash jarayonida til korpusining ahamiyati bir necha marta ortadi. Tadqiq predmeti ona tili bo‘lmagan (ikkinchi yoki xorijiy til hisoblangan) o‘qituvchi va o‘quvchi uchun ham korpus juda muhim va qulay vosita. O‘rni kelganda aytish lozimki, ilk rus tili korpuslari Rossiyada emas, Yevropada rus tili tadqiqotchilari tomonidan yaratilgan. [12]

Korpus materialining necha tilda berilishiga ko‘ra uning bir va ko‘p tilli turlari mavjud. Korpus mutaxassislarini (asosan, tarjimon) doim bir necha tilli korpus yaratish qiziqtirib kelgan. Korpus yaratishning ilk davridan boshlab ingliz, fin, fransuz, nemis, grek, norveg, ispan, shved va h. tillar uchun ikki tilli korpuslar paydo bo‘la boshlagan. Bunday korpus bitexts deb ham ataladi. Korpusni ikki tilli emas, balki uch, to‘rt va undan ortiq tilli qilishga hech qanday to‘siq yo’q. Mutaxassislar parallellik nuqtayi nazaridan korpusni bir, ikki va ko‘p tilli kabi turlarga ham bo‘lishadi. Bir tilli korpusda til varianti va shevalar bir-biriga qarama-qarshi qo‘yilsa, ikki va ko‘p tilli korpus bir mavzu doirasida turli tilda yozilgan

[11] Po‘latov A., Muhamedova S. Kompyuter lingvistikasi. – T., 2007. – B.43.

[12] Кутузов А.Б. Корпусная лингвистика. – М., 2005. С. 15-16.

matnlar majmuidan iborat bo'ladi. Masalan, ma`lum ilmiy muammo borasida turli davlatda turli tilda o'tkazilgan konferensiya materiallarini qamrab olishi mumkin. Ko'p tilli korpuslar, odatda, tarjimonlar tomonidan foydalaniladi. Ko'p tilli korpusning yana bir ko'rinishi original matn va tarjima matndan iborat bo'ladi. Korpusning ushbu turi qiyosiy chog'ishtirma tadqiqot olib borishda, tarjima nazariyasi hamda kompyuter tarjimasini o'rganishda juda muhim manba bo'lib xizmat qiladi. Ko'p tilli korpusning 2 turi mavjud:

1) bir-birining tarjimasi bo'lgan matnli korpus;
2) bir mavzuga oid ikki tildagi matnli korpus.

Birinchi tipdagi korpus "parallel korpus" (parallel corpora) deb nomlanib, ma'lum bir tarjimaning turli aspektini o'rganish uchun qo'llaniladi. Masalan, Kanada parlamenti yig'ini (ingliz/fransuz) matnlari korpusi mavjud. Parallel korpus o'z navbatida yana 2 turga – *moslashtirilgan* (aligned) va *moslashtirilmagan* (not aligned) korpusga ajraladi. *"Moslashtirilgan"* atamasi korpusda tarjima birliklari orasida bir-birini taqozo etuvchi aniq aloqa mavjudligini bildiradi. Bunday korpusning afzalligi u yoki bu gapning qanday tarjima qilinganini topishda qulaylik mavjudligida. Bu turdagi korpus tarjimon uchun ahamiyatli, chunki unda noyob resurs − "tarjima xotira"si (translation memory) mavjud. *"Moslashtirilmagan"* korpus ("qiyosiy" korpus)ning vazifasi matnni uning tarjimasi bilan moslashtirish, bir birlikning tarjimada qaysi birlikka to'g'ri kelishini ko'rsatib turishdan iborat. To'g'rilash avtomatik ravishda yoki qo'lda bajarilishi mumkin. Birinchi usul oson, lekin xatolari ko'p. Masalan, tarjima jarayonida sodda gap qo'shma gap shaklida berilishi mumkin. Bunday paytda qaysi qurilish original ekanligini aniqlash qiyin bo'ladi. Ko'p tilli "moslashtirilgan" korpus namunasi sifatida Yevropa Ittifoqining Acquis Communautaire ma'lumotlar bazasini keltirishimiz mumkin.

Ikkinchi xildagi korpus "tarjima korpusi" (translation corpora) deb atalib, ayni bir fikrning turli tildagi ifodasini o'rganish uchun muhim.

Parallel korpusning qimmati uning hajmi va tillarning miqdori bilan belgilanadi. Acquis Communautaire dunyodagi eng katta parallel korpus bo'lib, muhim jihati bu korpusdan foydalanishning bepulligi va malti-eston, sloven-fin kabi kam uchraydigan tillar juftligining mavjudligi bilan baholanadi. Ushbu korpuslardan quyidagi maqsadlarda foydalanish mumkin:

1) tipik tarjima usullari va transformatsiyani yuzaga keltirish;
2) avtomatik tarjima tizimi statistikasini o`rganish;
3) bir va ko`p tilli lug`atlar yaratish;
4) ma`lumotni saqlash va uzatish dasturlarini o`rganish va baholash;
5) tarjima to`g`riligini avtomatik tarzda tekshirish;
6) ekvivalent tanlash imkoni kengligi orqali tarjimon mehnatini osonlashtirish.

Dunyo tillarini qamrab olgan parallel korpusning ahamiyati turkiy tillar mushtarak korpusi tuzish masalasining nechog`liq dolzarbligini ko`rsatib turibdi. Turkiy tillar mushtarak korpusi matnshunoslik, qiyosiy tilshunoslik, tarjima nazariyasi, adabiyotshunoslik, qarindosh tillararo munosabatlar, til leksikasining boyish manbalarini o`rganish vositasi sifatida xizmat qilishi bilan ahamiyatli. Bunday korpusning yaratilishi turkiy tillar oilasiga mansub tillarning rivojlanishini ta`minlashi bilan birga, foydalanuvchilari kam sonli turkiy tillarni asrab qolish garovi hamdir. Turkiy tillarning mushtarak (parallel) korpusi turkiy tillarning mushtarak yodgorliklari – "Avesto", O'rxun-Enasoy obidalari, turkiy xalqlar eposlarini turkiyzabon xalq farzandlariga o`rgatishning eng zamonaviy ta'lim vositasi sifatida xizmat qilishi tabiiy. [13]

Jahon tillarining juda ko'pi mukammallik darajasi va matnni ilmiy qayta ishlash imkoniyati bilan farq qiluvchi o'z milliy korpusiga ega. Ingliz tilida yaratilgan Braun korpusi, Lankaster-Oslo/Bergen (LOB) korpusi, London-Lund korpusi, Leksikografik tadqiqotlar uchun Amerika meros korpusi, Lankaster ingliz tili so'zlashuv korpusi, diaxronik korpus sanalgan. Ingliz matnlarining

[13] Гарабик Р., Захаров В.П. Параллельный русско-словацкий корпус // Труды международной конференции «Корпусная лингвистика – 2006». – СПб., 2006. – С. 81-82.

Xelsinki korpusi, lingvodidaktik tadqiqotlar uchun Ingliz tili o‘rganuvchilarining xalqaro korpusi, ingliz tilidagi korpuslarning eng so‘nggi avlodi sifatida Ingliz tili banki, Britaniya milliy korpusi, Xalqaro ingliz tili korpusi, Amerika milliy korpusi kabi mashhur korpuslarning mavjudligi milliy va davlat tili taraqqiyotida milliy korpusning ahamiyati va o’rni nechog`lik muhimligini ko‘rsatadi. Dunyo tili sanalgan ingliz, ispan, xitoy, arab, frantsuz, rus, nemis tillarining milliy korpusidan tashqari, polyak, polyak-ukrain, chex, slovak, serb, xorvat, bosniya, bolgar, bolgar-rus, makedon, shotland, niderland, niderland-fransuz, shved, dat, norveg, island, farer, o`rta asr fransuz tili, italyan, portugal, rumin, litva, latish, grek, sharqiy arman, osetin, alban, hind, tsigan, xett, fin, ural tillari, eston, veps, venger, udmurt, gruzin, ingliz-gruzin, lezgin, turk, tatar, boshqird, qrim-tatar, qalmiq, bur¬yat, mo`g`ul, ivrit, amxar, yapon, qadimiy yapon, baman, esperanto tillari korpuslari mavjud. Dunyo kompyuter lingvistikasida milliy til korpusining mavjudligiga tilning yashovchanlik va kompyuter tiliga aylanish mezoni sifatida qaralmoqda.

2.2. O‘zbek tilining milliy korpusini yaratish mezonlari

Hozirgi o‘zbek tili boshqa tillar kabi o’z tarixiy taraqqiyot yo‘lida bir qancha o‘zgarishlarga uchradi. Taraqqiyot davridagi har bir bosqichning o‘ziga xos xususiyatini tadqiq etish maqsadida tilning taraqqiyot bosqichlari ajratildi hamda o‘rganildi, bu tadqiqotlarda davr tilining o‘ziga xos xususiyati ma’lum bir adib ijodi misolida yoritildi. Har bir davr ijodkorlari ijodiy merosini to‘liq qamrab oluvchi bir qancha lingvistik tadqiqot mavjud bo‘lsa-da, ular maxsus tizimlashtirilgan baza shaklini olmagan. Axborot texnologiyalari asrida bunday yaxlit tizimning o‘zbek (turkiy) yozma manbalari uchun mavjud bo‘lmasligi achinarli holat.

Bugungi kunda korpus lingvistikasi tilshunoslikda real hayotda tildan kompyuter va elektron korpuslar yordamida foyalanish bilan bog‘liq yangi yondashuv sifatida tushuniladi. Lingvistikaning sintaksis, semantika,ijtimoiy lingvistika kabi bo‘limlari til tuzilmasi yoki tildan foydalanishni bayon etish yoki

baholash maqsadiga ega bo‘lsa, korpus lingvistikasitilga oid tadqiqotlarning ko‘plab aspektlarida qo‘llash mumkin bo‘lgan keng tushuncha, metodologiya hisoblanadi.

Dunyodagi yirik tillarning milliy korpuslari yaratilgan va yaratilmoqda. Ulardan ko‘pchiligi tog‘risida atroflicha ma’lumot berildi. O‘zbek tilining milliy korpusini yaratish tilshunosligimiz oldidagi dolzarb vazifalardan biridir. O‘zbek tilining milliy korpusining yaratilishi shu bilan ahamiyatliki, buning natijasida tadqiqotchi ma’lumotlarni olish uchun behad ulkan axborot hajmiga ega bo‘ladi. Bu esa til birliklarining barcha lingvistik xususiyatlari, tilning taraqqiyoti, undagi o‘zgarishlar – yangilanish va eskirishlar, faollashish va passivlashishlar haqida tezkor, aniq va to‘liq ma’lumotni beradi, osonlik bilan turli tipdagi katta hajmli akademik lug‘atlar tuzish, matnlarga avtomatik ishlov berish imkoniyatini yaratadi.

Tilshunosligimiz tarixida alohida muallif asari til xususiyatini o‘rganishga bag‘ishlangan maxsus tadqiqotlar talaygina. Runiy va uyg‘ur yozuvi asosida yaratilgan yodgorliklarning til xususiyatiga bag‘ishlangan bir qancha asarlar maydonga keldi. Bu tadqiqot ishlarida qadimgi turkiy til xususiyatlarining ayrim morfologik, sintaktik va fonetik xususiyatlari tahlil qilindi. Runiy yozuvdagi yodgorliklar tilini V.V.Radlov, S.E.Malov, A.M.Sherbak, G`.Abdurahmonov, A.Rustamovlar atroflicha tadqiq etishdi. XI-XIV asr yozma yodgorliklari til xususiyati A.K.Borovkov ishlarida, “Devonu lug‘otit turk” asari tili A.Fitrat tadqiqotlarida, Yusuf xos Hojibning “Qutadg‘u bilig” asari til xususiyati V.V.Radlov, S.E.Malov, V.V.Bartold, E.E.Bertelslar, Ahmad Yugnakiyning “Hibatul-haqoyiq” asari S.E.Malov, E.E.Bertels, Q.Mahmudovlar tomonidan alohida o‘rganildi. O‘g‘uz dialekti belgilari ustun bo‘lgan “Qissayi Yusuf” dostoni M.Brokkelman, o‘g‘uz va qarluq-uyg‘ur dialektining qadimgi yodgorliklaridan hisoblangan “O‘g‘uznoma” afsonasi G.N.Potapin, A.M.Sherbak, XIV asr adabiy tili yodgorligi “Muhabbatnoma” asari til xususiyati ham A.M.Sherbak tomonidan tadqiq etilgan. XI-XIV asrlarda tuzilgan rasmiy hujjatlar uslubida o‘sha davr tiliga xos xususiyatlar o‘z ifodasini topgan. XI-XIV asrlarga oid rasmiy tilni bayon

etishda prof. S.E.Malov nashr qilgan va X-XIII asrlarga yorliqishli yuridik hujjatlar, O.D.Chexovich nashr etgan XIV asrga oid Buxoro hujjatlari hamda E.R.Tenishev nashr etgan XIII-XIV asrlarga oid xo‘jalik yozuvlari yordam beradi. Eski o‘zbek adabiy tili deb ataluvchi XIV-XIX asr tili xususiyati Atoiy, Sakkokiy, Lutfiy, Navoiy asarlari misolida tadqiq etilgan. Xususan, Alisher Navoiy asarlari tili alohida ahamiyatga ega bo‘lib, eski o‘zbek adabiy tili taraqqiyotida muhim o‘rni borligini A.K.Borovkov tadqiqotlari isbotlaydi. XVII-XIX asr adabiy tili xususiyatlari Abulg‘ozi Bahodirxon, Turdi Farog‘iy, Nishotiy, Munis, Gulxaniy, Maxmur asarlari tili, XX asr o’zbek adabiy tili leksikasi xususiyatlari Abdulla Qodiriy, G`afur G`ulom, Oybek, Abdulla Qahhor asarlari til xususiyatlari misolida tadqiq etilgan. Bu tadqiqotlarda muallif tili leksikasi atroflicha tavsiflangan. Bunday tadqiqotlar tilning rivojlanish jarayoni va o‘zgarishlarni: ma’no torayishi va kengayishi, istorizm, arxaizm, neologizmlar harakatini kuzatish uchun qulay sharoit yaratadi. Mana shunday xususiyat har bir muallif va davr tili uchun tadqiq etilsa ham, ular alohida-alohida monografiyalarda mavjud. Agar bu tadqiqotlar milliy korpusda yaxlit bir tizim asosida o‘z aksini topsa, istalgan leksemaning tarixiy taraqqiyot jarayonini turli muallif tili misolida yorqin tasavvur qilish imkoniyati paydo bo’ladi. Tilshunosligimiz oldida turgan endigi vazifa shu tadqiqotlarni birlashtirish asosida korpus yaratish. Bu vazifani amalga oshirish ikki bosqichni o‘z ichiga oladi:

1. Tilimiz tarixiy taraqqiyoti davrida til xususiyatlari tadqiq qilingan adiblar ijodiy merosi va tadqiqot natijasini korpus sifatida shakllantirish;

2. Shu paytgacha til xususiyati o‘rganilmagan yozuvchi va shoirlar asari korpusini tuzish.

Birinchi vazifa o‘rganilgan milliy-madaniy merosimizning elektron shaklda yaxlit sistemada (korpus shaklida) saqlanishi va undan keyinchalik turli ta`limiy va tadqiqiy maqsadda foydalanish imkonini yaratadi. Ikkinchi vazifani amalga oshirish til xususiyati o‘rganilmagan yozuvchi va shoirlar asariga bag‘ishlangan

turli filologik tadqiqotlar bajarish uchun elektron baza hamda milliy-madaniy meros va tilimizni elektron shaklda (korpus formatida) saqlash, foydalanish va kelgusi avlodga to‘liq yetkazishni ta`minlaydi.

Adabiy til rivojiga salmoqli hissa qo‘sha olgan buyuk milliy yozuvchilarning leksikasini asrab qolish, nafaqat asrash, balki butun ko‘lami bilan tavsiflash, til egalari uchun namuna vazifasini o‘taydigan model shakliga keltirish adabiy tilni asrash va rivojlantirish omili bo‘lib xizmat qiladi. Bu borada o‘z millati milliy va adabiy tili asoschisi hisoblangan butun dunyoga mashhur adiblar Shekspir, Dante, Gyote, Pushkinlarning adabiy merosi asosida qilingan ishlar (tuzilgan korpus) namuna sifatida xizmat qiladi. Bu siraga adabiy tilning eng optimal tashuvchisi sifatida tan olingan yozuvchilar Shandor Petefi, Genrik Ibsen, Adam Mitskevich, Karel Chapeklarning ishlarini kiritish ham mumkin. Korpus lingvistikasi ancha yutuqlarga erishgan bo‘lsa-da, mualliflik korpusi tuzish tajribasi oxirgi 5 yil ichidagina shakllandi va bu borada ancha natijalarga erishishga ulgurdi. Bunday korpus tuzish ma’lum davr adabiy tili, xususan, leksik tarkibi, uslubiy xususiyatini tadqiq etish uchun material vazifasini o‘tashdek qimmatli vazifani ado etadi. Yuqorida ta’kidlanganidek, mualliflik korpusini tuzishning bir qancha afzalliklari bor. Mualliflik korpusi, uning o‘ziga xos xususiyati, tuzish tamoyili kabi muammolar ushbu bo‘limning tadqiq predmeti hisoblanadi. Mualliflik korpusi tarkibiy qismi, interfeysi, razmetka masalasi korpus tuzishning asosiy muammosi hisoblanganligi sababli avval Internet tarmog‘ida mavjud mualliflik korpuslarining shu jihatiga e’tibor qaratish maqsadga muvofiq.

Korpus uchun material tanlash: muammo va yechim**.** Milliy korpusi mavjud tillarning mualliflik korpusini tuzish hech qanday qiyinchilik tug‘dirmaydi. Chunki matn (muallif asari) elektron shaklda internet saytida yoki korpus tarkibida mavjud. Mualliflik huquqi asosida bu asarlar tarmoqda turgani uchun ularni korpusga xom ashyo sifatida olishga monelik yo‘q. Milliy korpusi mavjud bo‘lmagan tillarda korpus yaratishda esa bu borada biroz muammoga duch kelish tabiiy hol. O‘zbek tili milliy korpusi bo‘lmasa ham, ZiyoNET tarmog‘ida mumtoz hamda zamonaviy shoir va yozuvchilarimiz asarlarining anchasi elektron shaklda

joylashtirilgan. Shunga asosan, mualliflik korpusi tuzishda ikki manbaga asoslanish mumkin:

1. Har bir muallifning nashr etilgan mukammal asarlari to‘plami.
2. Internet tarmog‘idagi elektron fayllar.[14]

Birinchi manbani elektron shaklga aylantirish (skanerlash, matnni sun’iy intellekt tushunadigan formatga keltirish) orqali material sifatida ishlatsak, ikkinchi manbadan nisbatan tayyor holda foydalanishimiz mumkin. Ikkala holda ham matnning elektron shakli olingach, uni texnik qayta ishlash – tokenizatsiya, lemmatizatsiya, sintaktik razmetkalash ehtiyoji tug‘iladi. Texnik ishlov berishdan oldin matn korpus uchun tayyorlanadi, chunki unda nolingvistik birlik ham uchraydi. Korpus matnining asosiy belgisi unda nolingvistik birlikning (jadval, rasm, grafik chizma) bo’lmasligidir. Shundan so‘nggina razmetkalash bosqichiga o‘tish mumkin. Razmetkalash avtomatik va yarimavtomatik rejimda bajariladi. Muallif qo‘llagan neologizm, boshqa alifbodan yozilgan so‘zshakl lemmatizatsiyasiga alohida e’tibor qaratiladi. Korpus tayyor holga kelgach, tabiiy ravishda muallif asarlari chastotali lug‘ati tayyor bo‘ladi. Chunki lemmatizatsiya jarayonida so‘zshakl hamda leksema (lemma) miqdori aniqlanadi. Masalan, Chexov asarlari korpusi 36 153 lemma yoki leksemani qamrab oladi. Ushbu lemmalar 1 381 000 qo‘llanish holati (120 000 so‘zshaklda)ni tashkil etgan. So‘z qo‘llash holatiga qarab, gapning o‘rtacha uzunligi (nechta so‘zdan iboratligi) ham aniqlanadi. Leksema qo‘llanish chastotasi asar yozilish yili, janri, gap uzunligi asosida ham hisoblanishi mumkin. Bu esa foydalanuvchiga hozirgi adabiy til va muallif davri adabiy tili leksikasi chastotasini qiyoslash va xulosa chiqarish imkonini beradi. Bunday qiyosiy tahlil milliy korpus asosida bajarilishi ham mumkin. Demak, bunday korpus tuzishdan yana bir maqsad muallif asarlarining turli lug`atini yaratish. Shu bilan birga, tarix nuqtayi nazardan katta davr ichidagi tilning tarixiy-madaniy rivojlanish va o‘zgarishini ham o‘rganish mumkin.

O‘zbek tilining milliy korpusini yaratish uchun, avvalo, kerakli materiallar to‘planadi. Korpusni shakllantirishda o‘zbek tilida yaratilgan veb-saytlar,

[14] Mengliyev B va boshqalar. O‘zbek tilining milliy korpusi//Ma’rifat, - Toshkent, 2018.

shuningdek, kutubxonalardan olingan elektron kitoblar va maqolalar asosiy manba sifatida xizmat qiladi.

Korpus ma’lumotlari O‘zbekiston Respublikasi qonunlariga muvofiq litsenziya asosida tarqatiladi. Ularda materialning qaysi manbadan olinganligi qat’iy ko‘rsatiladi.

O‘zbek tilining milliy korpusi o‘zbek tilidagi matnlarning electron shakldagi axborot-ma’lumot tizimi hisoblanadi. O‘zbek tilining milliy korpusi saytga (masalan, http://uzbekcorpora/uz/) joylashtiriladi. Korpus o‘zbek tili bilan bog‘liq masalalar bilan qiziquvchi va undan foydalanuvchi – tilshunoslar, tarjimon va tarjimashunoslar, til o‘rganuvchilar, o‘quvchilar va talabalar, o‘zbek tilini o‘rganayotgan chet elliklar uchun mo‘ljallanadi.

O‘zbek tilining milliy korpusi zamonaviy korpuslarga qo‘ladigan barcha talablarga javob berishi va quyidagi xususiyatlarga ega bo‘lishi kerak:

1) so‘z hajmi;
2) o‘zbek tilining barcha foydalanish sohalariga yorliqishli matnlar janrining xilma-xilligi:
   - badiiy uslubdagi matnlar (XX asr boshidan to bugungi kunga qadar yaratilgan adabiy matnlar);
   - publitsistik uslubdagi matnlar (keying o‘n yillikda internetda joylashtirilgan maqolalar);
   - rasmiy uslubdagi matnlar (2010-2014 yillarda e’lon qilingan farmoyishlar, qarorlar, buyruqlar va h.k rasmiy hujjatlar);
   - ilmiy uslubdagi matnlar (turli sohalarda yaratilgan ilmiy tadqiqotlar, monografiyalar va h.k.);
   - so‘zlashuv tilidagi matnlar (2010 yildan beri yaratilgan mashhur blog-postlar);
3) asosiy ijtimoiy parametrlar (yoshi, ma’lumoti darajasi, tilni bilish darajasi, kasbi, nutq madaniyati turlari) bo‘yicha turfa mualliflar tarkibi;
4) turli davrlarga yorliqishli matnlarning mavjudligi.

Ost korpuslarni annotatsiyalashda sintaktik va leksik yorliqsetlar ishlab chiqariladi. Sintaktik yorliqsetda sodda gap С, bosh gap С, ergash gap СБАР, САРҚ, ega НП, WҲНП, aniqlovchi АДЖП, hol ПП, WҲП, АДВП WҲАДВП va nolga teng, bosh gap bo‘lagi X kabi shartli belgilar asosida qolipga solinadi.

O‘zbek tili agglutinativ tillar guruhiga xosligi uchun so‘z shakllari so‘z o‘zagiga ketma-ketlikda birikkan morfemalar qatoridan tashkil topadi. Morfemalar o‘z navbatida turli grammatik xususiyatlar (shaxs, son, kelishik va h.k.) bilan xarakterlanadi va o‘zida muhim kontekst informatsiyasini tashiydi, buni hisobga olmagan leksik tahlil to‘liq bo‘lmaydi. Shunga ko’ra, leksik yorliqsetda dastlab grammatik xususiyatlarni ishlab chiqish kerak bo‘ladi. Leksik yorliqsetdagi grammatik xususiyatlarni quyidagicha belgilash maqsadga muvofiq: son Н 2, egalik С 10, shaxs П 8, kelishik С 7, bo‘lishsizlik Г 2, zamon Т 3, tuslash М 4, mayl В 5.

Korpusni shartli ravishda ikki guruhga ajratish mumkin:

- zamonaviy;
- diaxronik.

Zamonaviy matnlar korpusiga yaratilish davri muayyan yillarni o‘z ichiga oluvchi matnlar kiritiladi. Korpus ushbu qismining asosiy hajmi so‘z ishlatmalarini o‘z ichiga oladi. Diaxronik qism ma’lum miqdordagi so‘zlik hajmiga ega bo‘lib, u muayyan asrga yorliqishli matnlarni o‘z ichiga oladi. O‘zbek tilining milliy korpusi hajmi chastotali til ko‘rinishlarining variativligi va o‘zgaruvchanligini o‘rganish, shuningdek, quyidagi yo‘nalishlar bo‘yicha ishonchli natijalarni qo‘lga kiritish imkonini yaratishi lozim:

1) so‘z turkumlarining morfologik variantlari va ularning evolyutsiyasini o‘rganish;
2) so‘z yasash variantlari va ular bilan bog‘liq so‘z yasalish modellari hamda vositalari samaradorligi muammolarini tadqiq etish;

3) boshqarish, moslashtirish , biriktirish variantlarining o‘zgarishini tadqiq etish;
4) akseptologik variantlar va o‘zbek tilining aksept tizimidagi o‘zgarishlarni tadqiq qilish;
5) leksik variativlik, xususan, sinonimik qatorlar va tematik guruhlar, tarkib, shuningdek, ulardagi semantik nisbatlarning o‘zgarishini o‘rganish.

O‘zbek tilining milliy korpusiga quyidagi korpusostilar ham kiritiladi:

- chuqur taqrizlangan korpus – undagi har bir gap uchun to‘liq morfologik va sintaktik qurilma yaratiladi;
- matnlarning parallel o‘zbekcha-inglizcha korpusi unda muayyan o‘zbekcha yoki inglizcha so‘z yoxud so‘z birikmasining barcha tarjimalarini topish mumkin;
- dialektal matnlar korpusi – bunda O‘zbekistonning turli mintaqalariga yorliqishli dialektal nutqi ularning grammatik spetsifikatsiyasini saqlangan holdagi yozuvlari kiritiladi, dialektal morfologiya hisobga olingan maxsus qidiruv e’tiborga olinadi;
- poetik matnlar korpusi – unda nafaqat leksik va grammatik belgilar, balki she’r uchun o‘ziga xos bo‘lgan belgilar (epigrammlar va she’rlarning muayyan o‘lchamlari, qofiyalanishlari va boshqalar) bo‘yicha ham qidirish imkoniyati mavjud bo‘ladi;
- o‘zbek tilini o‘rganish korpusi annotatsiyasi o‘zbek tilini o‘zgarishning maktab dasturiga yo‘naltirilgan, har qanday omonimiyadan xoli korpusi;
- og‘zaki nutq korpusi ommaviy va xususiy og‘zaki nutqning magnitafon yozuvlari va kinofilmlari transkripsiyalari rasshivrovkasini o‘z ichiga oladi.

Mavjud milliy korpuslarni kuzatish asosida korpusning tuzilishi va tarkibini o`rganar ekanmiz, korpus interfeysi, qidiruv tizimi va matnlar bazasi uning eng asosiy tarkibiy qismi, degan xulosaga kelamiz. Rus tili milliy korpusi joylashgan www.ruscorpora.ru saytining birinchi sahifasida korpus va uning tuzuvchilari

haqida asosiy ma`lumot, o‘ng tomondagi menyuda istalgan sahifaga o`tish imkoniyati mavjud. Bu korpus menyusi to`rt qismdan iborat. Bosh sahifa, saytning qidiruv resursi, matn haqida unga biriktirilgan qo‘shimcha ma‘lumot ilovasi, korpus birliklariga izoh yozish prinsiplari, oxirgi blok korpus tuzuvchilari jamoasi, foydalanilgan dastur, matnlarning mualliflik huquqi haqida to‘liq ma’lumotlar bazasidan iborat.

Albatta, korpusning tuzilishi va tarkibi tilning xususiyatlari, ijtimoiy talab va boshqa jihatlarga ko‘ra turlicha bo‘lishi mumkin. Korpuslar uchun yagona va o‘zgarmas andoza belgilanmaydi. Masalan, tuzilajak o`zbek tili korpuslari jamiyatimiz talablaridan kelib chiqqan holda o‘ziga xos bo‘lishi mumkin. Korpusdagi milliy so‘zi nafaqat tilning, balki korpus tuzilishi va tarkibining ham o‘ziga xosligini anglatadi.

II. O‘ZBEK TILINING MILLIY KORPUSI VA KORPUS YARATISH TAMOYILLARI

2.1. Korpus yaratishda lingvistik belgi yorliqlari

Korpuslar o‘zining xususiyatlari, tadqiqot uchun bazaviy darajasi katta-kichikligiga ko‘ra oddiy va tavsiflangan til korpuslariga ajraladi. Demak, biz tavsiflangan korpus qanday bo‘lishini aniqlashtirib olishimiz kerak. Lingvistik tavsiflangan yoki unga belgi qo‘yilgan degani bu korpusga matndan tashqari, matnga aloqasi bo‘lmagan, lekin qandaydir bir ma’lumot beruvchi manba sifatida izohlanishi mumkin. Buni quyidagicha tushunish mumkin:

I will use Google before asking dumb questions.

Belgilaymiz:

*I (**pronoun**) will (**verb**) use (**verb**) Google(**noun**) before(**preposition**) asking (**verb**) dumb(**adjective**) questions(**noun**).*

Asosan, buni avtomatik analiz korpus ishini osonlashtirishda foydalanamiz. Til korpusiga birinchi marta razmetka qo‘yilayotganda, istalgan belgini qo‘yishimiz mumkin, faqat belgi va so‘z orasida qandaydir bog‘liqlik bo‘lishi kerak. Shunda shu so‘zga o‘xshash so‘zlarni topishda ham qiynalmaymiz.

80-yillarda elektron matnlarga standart belgilar kiritildi., ya’ni qabul qilindi. Avvalo, ushbu belgi tipografiya(bosmaxona) sohasi uchun mo‘ljallangan bo‘lsa-da, keyinchalik barcha sohalarda keng qo‘llanila boshlandi. SGML (Standard Generalized Markup Language) nomi ostida bu hodisa keng tarqaldi. Biz u orqali har xil shakldagi matnlarni tahlil qilish, tahrir qilish hamda o‘zgartirish imkoniyatiga ega bo‘ldik. [15]

SGML yorliqlar g‘oyasiga asoslanadi. Yorliqlar – bu matndagi o‘ziga xos belgilar tizimi bo‘lib, mazkur tekst haqida informatsiya beradi. Har qanday holatda har qanday matnlar uchun razmetka yorliqlaridan foydalanishimiz mumkin.

[15] Кутузов А.Б. Корпусная лингвистика. – М., 2005. С. 155-156.

Til belgilari bo‘lgan SGML bu til konstruktorlari sanaladi. U o‘zining ilk holatidagi ko‘rinishida va ishlatilishida murakkabligi sababli uning bazasida **HTML** va **XML** yaratildi. Bugungi kunda butun dunyo o‘rgimchak to‘rida bu dasturlardan keng foydalanilmoqda. Ushbu dastur orqali biz matnlarni barcha parametrlarga ko‘ra: tagiga chizish, sitatalarni aniqlash, boshqa tildan kirib kelgan so‘zlarni aniqlash, ro‘yxat tuzish va shu kabi tadqiqiy vazifalarni bajarishimiz mumkin.

O‘zbek tilida semantik razmetka yorliqlarining lingvistik modellari

Ot so‘z turkumi morfologik va semantik belgi yorliqlari

Atoqli ot ma’noviy guruhlari [atoqli ot] = [at. ot.]

1. [shaxs nomi] = [sh.n.]
2. [geografik nom] = [geog.n.]
3. [tashkilot, muassasa, korxona nomi] =[TMK n.]
4. [samoviy yoritgichlar nomi] = [SYO n.]
5. [tarixiy sana yoki bayram nomi] = [TS n.]
6. [hayvon nomi] = [h.n.]
7. [mahsulot nomi] = [m.n.]
8. [ilohiy tushuncha nomi] = [iloh.n.]

Turdosh ot ma’noviy guruhlari [turdosh ot] = [tur. ot]

[mavhum ot] = [mavh.o.]

[aniq ot] = [an.o.]

Aniq ot LMTlari oid yorliqlar tizimi

1. [modda nomi] = [m.n.]
2. [shaxs nomi] = [sh. n.]

1) [shaxsni qarindoshlik jihatdan tavsiflovchi LMG] = [qar. LMG]
2) [shaxsni yosh jihatdan tavsiflovchi LMG] = [yosh LMG]
3) [shaxsni kasb jihatdan tavsiflovchi LMG] = [kasb LMG]
4) [shaxsni jins jihatdan tavsiflovchi LMG] = [jins LMG]

5) [shaxsni istiqomat jihatdan tavsiflovchi LMG] = [istiqomat LMG]

6) [shaxsni mansab-unvon jihatdan tavsiflovchi LMG] = [m.u. LMG]

7) [shaxsni ijtimoiy holat jihatdan tavsiflovchi LMG] = [ijt. h. LMG]

3. [buyum nomi] = [b.n.]

4. [o‘simlik nomi] = [o‘.n.]

5. [o‘rin-joy nomi] = [o‘.j.n.]

6. [miqdor nomi] = [miq.n.]

7. [tashkilot va muassasa nomi] = [t.m.n.]

8. [vaqt-payt nomi] = [v.p.n.]

9. [faoliyat-jarayon nomi] = [f.j.n.]

3. Modda nomi LMT tarkibiy qismi tashkil etuvchi birliklar yorlig‘i

1. Yonuvchi foydali qazilmalar = [3.1.]

2. Rudali foydali qazilmalar = [3.2.]

a) qora metallar = [3.a]

b) rangli metallar = [3.b]

3. noruda foydali qazilmalar = [3.3.]

4. **Olam tushunchasi tarkibiy qismlari yorlig‘i**

1. Jonsiz tabiat. Buyumlar = [4.1.]

2. O‘simliklar va hayvonot dunyosi. Inson (tirik organizm) = [4.2.]

3. Fazo. Fazoviy holat. Shakl = [4.3.]

4. O‘lcham. Munosabatlar. Sababiyat = [4.4.]

5. Vaqt = [4.5.]

6. Dunyo. Rang. Tovush. Temperatura. Og`irlik = [4.6.]

7. Holat. Hid bilish. Maza-ta`m (zohiriy va botiniy qabul qilish) = [4.7.]

8. Harakat = [4.8.]

9. Istak va xohish = [4.9.]

10. Sezish = [4.10.]

11. Tuyg`u. Jazava (qayajon). Xarakter qirralari = [4.11.]

12. Tafakkur = [4.12.]

13. Belgi. Axborot. Til = [4.13.]

14. Yozuv. Bilim (fan) = [4.14.]

15. San`at = [4.15.]

16. Jamiyat va ijtimoiy munosabatlar = [4.16.]

17. Asbob-uskunalar. Texnika = [4.17.]

18. Qishloq xo‘jaligi = [4.18.]

19. Huquq. Etika = [4.19.]

20. Din. G`ayritabiiy hodisalar = [4.20.]

5. Ilohiyot va osmon jismlari rukni yorliqlari

1. Olloh va unga bog‘liq bo‘lgan tushunchani ifodalovchi nomlar mikromaydoni = [5.1.]

2. Diniy tushunchani ifodalovchi so‘zlarning semantik maydoni = [5.2.]

3. Kosmonimlar = [5.3.]

4. Ob-havo va shamol bilan bog`liq tushunchani anglatuvchi so`zlar semantik maydoni = [5.4.]

5. Islom dini bilan bog`liq tushunchani ifodalovchi mikromaydon = [5.5.]

6. Boshqa dinlar bilan bog`liq tushunchalar maydoni = [5.6.]

6. “O`simliklar dunyosi” (fitonimlar) maydoni yorliqlari

1. Suv o`tlari = [6.1.]

2. Zamburug‘lar = [6.2.]

3. Yovvoyi o‘tlar (ko‘katlar) = [6.3.]

4. Madaniy gullar (ko‘katlar) = [6.4.]

5. Zararli o‘simliklar = [6.5.]

6. Dorivor o‘simliklar = [6.6.]

7. Boshoqli o‘simliklar = [6.7.]

8. Qora ekinlar = [6.8.]

9. Poliz ekinlari = [6.9.]

10. Ozuqabop o‘simliklar = [6.10.]

11. Sanoat ekinlari = [6.11.]

12. Cho‘l, adir, tog‘ o‘simliklari = [6.12.]

13. O‘rmon o‘simliklari = [6.13.]

14. Butalar = [6.14.]

15. Yovvoyi daraxtlar = [6.15.]

16. Mevali daraxtlar = [6.16.]

7. Hayvonot dunyosi maydoni yorlig‘lari

1. Bir hujayrali hayvonlar = [7.1.]

2. Hasharotlar = [7.2.]

3. Parazit hasharotlar = [7.3.]

4. Sudraluvchilar = [7.4.]

5. Baliklar = [7.5.]

6. Amfibiyalar = [7.6.]

7. Qisqichbaqasimonlar = [7.7.]

8. Suv hayvonlari = [7.8.]

9. Xonaki qushlar = [7.9.]

10. Yovvoyi qushlar = [7.10.]

11. Uy hayvonlari = [7.11.]

12. Yovvoyi hayvonlar = [7.12.]

13. Yirtqich hayvonlar = [7.13.]

14. Afsonaviy hayvonlar = [7.14.]

8. Buyum nomi tarkibini bildiruvchi birliklar yorlig‘i

1. Mexanizm va jihoz = [8.1.]

2. Transport vositasi = [8.2.]

3. Qurol-yarog` = [8.3.]

4. Musiqa asbobi = [8.4.]

5. Mebel = [8.5.]

6. Idish-tovoq = [8.6.]

7. Kiyim-kechak = [8.7.]

8. Oziq-ovqat va ichimlik = [8.8.]

9. Kiyim-kechak nomlari mazmun guruhi yorliqlari

1. Bosh kiyimi = [9.1.]

2. Ustki kiyim = [9.2.]

3. Ichki kiyim = [9.3.]

4. Oyoq kiyimi = [9.4.]

5. O‘ragichlar = [9.5.]

10. Faqat qism nomini birdiruvchi so‘zlar yorlig‘i

1. Inson tana a`zosi va organ = [10.1.]

2. Hayvon tana a`zosi va organ = [10.2.]

3. O‘simlik qismlari = [10.3.]

4. Bino va inshoot qismi = [10.4.]

5. Qurilma qismi = [10.5.]

6. Asbob qismi = [10.6.]

7. Mexanizm va jihoz qismi = [10.7.]

8. Transport vositasi qismi = [10.8.]

9. Qurol-yarog‘ qismi = [10.9.]

10. Musiqa asbobi qismi = [10.10.]

11. Mebel qismi = [10.11.]

12. Idish-tovoq qismi = [10.12.]

13. Kiyim-kechak va oyoq kiyimi = [10.13.]

Fe`l so`z turkumining morfologik va semantik razmetka LMGlari yorlig‘i

1) harakat fe’li = [har. f.]

2) nutq fe’li = [nutq f.]

3) holat fe’li = [holat f.]

4) natijali faoliyat fe’li = [nat. f. f.]

5) tafakkur fe’li = [tafakkur f.]

6) munosabat fe’li = [munosabat f.]

13. Nutq fe`llari tarkibiy LMGlari yorlig‘i

1) demoq fe’li = [13.1]

2) gapiruv fe’llari = [13.2]

3) ifodalov fe’llar = [13.3]

4) talaffuz fe’llari = [13.4]

5) sub`ektiv hukm fe’llari = [13.5]

6) nutqiy da’vat fe’llari = [13.6]

14. Fe’l leksemalarning mavzu to‘dalari yorlig‘i

1. Tirik organizmlarning barchasi uchun umumiy bo‘lgan LMT = [14.1]

2. Insonga xos harakat va holatni ifodalovchi LMT = [14.2]

3. Hayvonlarga xos harakat va holatni ifodalovchi LMT = [14.3]

a) xonaki hayvonlarga xos HHIL = [14.3. a]

b) yovvoyi hayvonlarga xos HHIL = [14.3. b]

4. Narsa va jismlarning harakat va holatini ifodalovchi LMT = [14.4]

a) daraxtlar va o‘simliklarga xos HHIL = [14.4.a]

b) tabiiy jismlarga xos HHIL = [14.4.b]

5. Faqat insonga xos bo‘lgan HHIL = [14.5]

1) Inson hayotining umumiy qirralariga xos HHILning MT = [14.5.1.]

2) Insonning tana a’zolari HHILning MT = [14.5.2.]

a) ko`z HHIL = [14.5.a.]

b) qo`l HHIL = [14.5.b.]

c) oyoq HHIL = [14.5.c.]

d) til HHIL = [14.5.d.]

e) tish HHIL = [14.5.e.]

f) lab HHIL = [14.5.f.]

g) qosh HHIL = [14.5.g.]

h) peshona HHIL = [14.5.h.]

i) quloq HHIL = [14.5.i.]

j) yurak HHIL = [14.5.j.]

k) qorin HHIL = [14.5.k.]

3) Inson hayotining ijtimoiy qirralarini o‘zida mujassam etgan fe’l-leksemalar = [14.5.3.]

a) tafakkur bilan bog‘liq leksema = [14.5.a]

b) dala ishlari bilan bog‘liq leksema = [14.5.b]

c) texnika bilan bog‘liq leksema = [14.5.c]

d) qurilish ishlari bilan bog‘liq leksema = [14.5.d]

e) emoq-ichmoq bilan bog‘liq leksema = [14.5.e]

f) kiyim-kechak bilan bog‘liq leksema = [14.5.f]

g) uy hayvonlarini boqish bilan bog‘liq leksema = [14.5.g]

6. Yagona subyekt bilan bog‘liq HHIL = [14.6]

7. Dialogik nutq elementlarini o`zida ifoda etgan leksemalar = [14.7]

8. Bir nechta subyektga yorliqishli bo`lgan HHIL = [14.8]

Ravish so‘z turkumining semantik razmetkasiga oid yorliqlar

Morfologik razmetkadagi yorliqlar tizimi

1) payt ravishi = [p. rav.]

2) o`rin ravishi = [o’. rav.]

3) holat ravishi = [h. rav.]

4) miqdor-daraja ravishi = [m.d. rav.]

5) maqsad ravishi = [maq. rav.]

6) sabab ravishi = [sab. rav.]

Semantik razmetka yorliqlari

1. Holat ravishi (HR):

1) yurish fe’llari bilan birikib keladigan HR = [15.1]

2) nutq fe’llari bilan birikib keladigan HR = [15.2]

3) tafakkur fe’llari bilan birikib keladigan HR = [15.3]

4) natijali faoliyat fe’llari bilan birikib keladigan HR = [15.4]

2. Payt ravishi:

1) payt mazmunining boshlanish nuqtasini ifodalovchi ravish = [16.1]

2) payt mazmunining davomiyligini ifodalovchi ravish = [16.2]

3) payt mazmunining izchilligini ifodalovchi ravish = [16.3]

4) payt mazmunining uzluksizligini ifodalovchi ravish = [16.4]

5) payt mazmunining chegaralanganligi yoki galma-galligini ifodalovchi ravish = [16.5]

3. Miqdor-daraja ravishi

1) kuchaytiruv ravishi = [17.1]

2) kuchsizlantiruvchi ravish = [17.2]

3) me`yoriy darajadagi ravish = [17.3]

2.2. Dunyodagi mavjud til korpuslarining yaratilish tamoyillari

Korpus lingvistikasi kompyuter lingvistikasining bir bo‘limi bo‘lib, lingvistik korpuslar(matnlar)ni tuzish va ularni kompyuter texnologiyasiga asoslanib kreativ ta’min yaratishni o‘rganadi. Lingvistika asosiga qurilgan ushbu yangi fan rivojlangan maxsus strukturaga ega bo‘lgan bo‘lib, mukammallashgan, tartibga solingan til qoidalari yoki lingvistik vazifani bajarishni o‘rganadi. Bugungi kunda “korpus” termini turli ma’nolarda qo‘llaniladi.

Korpus – matnni reprezentativ tuzish bo‘lib, odatda, mashina o‘qiy oladigan formatda va informatsiyaning tarkibiga vaziyatni soladigan (joylaydigan) matn turidir. Tilshunoslik yoki til nuqtayi nazaridan matnlar korpusining tanasi katta, mashinadan o'qiladigan shaklda ko‘rinadigan, yagona, tizimli, belgili, filologik jihatdan malakali til majmuasi hamda tilshunoslikka oid muayyan ma'lumotlar beruvchi baza sifatida tushuniladi. Hozirda “korpus” tushunchasini anglatishda ko'plab ta'riflar mavjud. Masalan, V.V.Rikov korpusni mantiqiy fikr, mantiqiy analiz sifatida o‘rganadi. E. Finegan: “Korpus - bu odatda, kompyuter o‘qiy oladigan formatda bo‘lgan va bizga matn ishlab chiqilgan vaziyat, informatsiya beruvchi, muallif, adresat yoki auditoriya haqidagi ma’lumotni o'z ichiga olgan matnlar to‘plamidir”, - deydi. [42] Turli umumiy ma’lumot beruvchi ijtimoiy saytlar korpusni statistik analiz hamda farazlar tekshiruvi asosida anglangan sohalarda uchrovchi til qoidalari va hodisalarini asoslay oladigan katta hajmdagi, tizimli matnlar to‘plami (endilikda odatiy elektron shaklda) sifatida ta’riflaydi. [62]. T.Mkeneri va E.Vilson quyidagi fikrlarni bildiradi: : “Korpus aniq til mezonlariga

muvofiq tanlangan, til namunasi sifatida foydalanish mumkin bo‘lgan tilshunoslik sohasidir”. [16]

Korpuslarni yaratish va o‘rganish quyidagi protsedurada ko‘rinadi:

1. Til mavjudligini katta planda yoritib uning ich-ichiga kirish;
2. Maxsus kontekstni formaga kiritib, har tomonlama va obyektiv o‘ranish;
3. Har xil g‘oya va maqsadlarni kompleks holda o‘rganish.

“Korpus matni” yoki “matnlar korpusi” boshqaruv tizimi, sistema ham sanalib, oxirgi yillarda bu sohani korpus menedjeri deb ham atashyapti. Bu maxsus qidiruv sistemasi bo‘lib, ma’lumotlarni qidirish uchun mo‘ljallangan. Til va nutq hodisalarini tez va qulay holda ajratishni ko‘zlaydi. Korpuslar turli savollar, yuborilgan so‘roqlar uchun statistik asos olish uchun qo‘l keladi. Shuni yodda tutish kerakki, korpus faqatgina til bo‘laklari asosida ish ko’radi. Korpuslar asosida so‘z, leksema, grammatik kayorliqoriyalar, chastotalar o‘zgarishi, turli vaqt natijasida o‘zgarishga uchragan kontekstlar ko‘riladi va tahlil etiladi. Lingvistik korpus haqida gapirar ekanmiz, biz avvalo, uning mohiyatiga yetib borishimiz kerak. Ushbu tushuncha ko‘plab ta’riflar orqali talqin etilgan. Lekin ularning barchasi bitta nuqtaga borib taqaladi. Uarni ko‘rib o‘tamiz:

- korpus bu bir timsolda, yagona umumiy belgilar asosida tashkil etilgan matn yoki matnlar massividan iborat yirik birlikdir;
- korpus bu lingvistik to‘plam bo‘lib, matnli, umuman olganda, yozuv shaklidagi tahlil uchun hammabop ma’lumotlardir; qolib ketgan

Korpus quyidagi turdagi matnlardan iborat bo‘lishi mumkin :

- ma’lum bir yozuvchining matnlari;
- ma’lum muddatdagi o‘n yillik yoki yuz yillik (asriy) matnlar;
- ma’lum mavzudagi zamonaviy matnlar;

[16] McEnery T, Wilson A. Corpus Linguistics. Edinburgh: Edinburgh University Press, 2nd edition, 2001.

- xalq tilini va jamiyat xususiyatlarini konkret tarzda aks ettirib turuvchi zamonaviy matnlar;

Korpus haqidagi bir talqinda uning yozma va og‘zaki bo‘lishi mumkinligi keltiriladi. Demak, matnlar to‘plami bo‘lgan korpus tahlil etilishi, uning yordamida ishlatilayotgan so‘zning kontekstdagi shakllarini, variantlarini va ketma-ketligini, tillarni va so‘zlarni bir-biriga solishtirilishi mumkin.

Hozirda korpusga oid tadqiqotlarga ko‘pchilik qiziqish bildirmoqda. Kompyuter dasturlarining rivojlanishi esa bu holatning taraqqiyotiga olib keladi. Chunki kompyuter yordamida juda katta hajmdagi matnlarni ishlab chiqish mumkin.

Vladimir Rikovning jadvalida korpus lingvistikasi va an’anaviy tilshunoslik o‘rtasidagi farqlar yaqqol ko‘rsatilgan. Ular quyidagilar:

Korpus lingvistikasi	**An’anaviy tilshunoslik**
Asosiy e’tiborini nutqni o‘rganishga qaratadi	*Asosiy e’tiborini tilni o‘rganishga qaratadi*
Asosiy maqsadi tilning nutqda qanday bayon etilganini, aks etganini ko‘rsatish	*Asosiy maqsadi tilni tavsiflash va tushuntirish orqali mohiyatiga borish*
O‘z tadqiqotida matnlar korpusi ma’lumotlariga tayanadi	*O‘z tadqiqotini nutqdagi hodisalar bilan boyigan nazariyadagi faktlar asosida olib boradi*
O‘z tadqiqotida kvantitativ (miqdor) metodlaridan foydalanishni afzal ko’radi	*O‘z tadqiqotida kvalitativ (sifat) metodlaridan foydalanishni afzal ko’radi*
Qisman an’anaviy, asosan, empirik (amaliy) metodlarga asoslanadi	*Qisman an’anaviy, asosan, fahmiy (nazariy) metodlarga asoslanadi*
Konkret til grammatikasida tuziladi	*Tillarning umumiy xususiyatlarini o‘rganadi*
Asosiy e’tiborini shaklni ajratishga qaratadi	*Asosiy e’tiborini nafaqat shaklga, balki mazmunga ham qaratadi*

Tekst (matn)ni global, ya'ni umumiy planda ko'rib chiqadi	*Tekst (matn)ni xususiy planda, tor doirada ko'rib chiqadi*

Ko'pchilik lingvistlar o'z tadqiqotlarida katta hajmdagi korpuslar bo'lgan British National Corpus hamda Cobuild Projectga murajaat qilishadi. Shunga qaramasdan har bir matn milliy ruhni aks ettirgani uchun hamda uning ustida bir til misolida ishlash shart bo'lganligidan o'z shaxsiy korpuslarini boshdan o'zlari yaratishadi. Biror tilning milliy korpusini yaratish uchun har qanday lingvist yoki ular jamoasi quyidagi tamoyillardan fordalanadi:

Rejalashtirish. Lingvistik korpus – bu matnlar yig'masi bo'lib, ularning bir joyga yig'ilishi asosida mantiq yotadi. Ushbu matnlarda mantiqiy fikr, mantiqiy prinsiplar mavjudligidan ular mushtarak holda tadqiqot obyekti bo'la oladi.

Korpus turi va ularning sistemasi korpusning nima maqsadda qo'llanilishiga bog'liq. Masalan, bizga qaysidir jurnaldagi reklama, ya'ni tijorat matnlari kerak. Biz dastur tizimi yordamida jurnaldagi 2000-yildan to bugungi kunga qadar chop etilgan reklamaga oid matnlarni taqqoslab ko'rish imkoniga egamiz.

Demak, biz tilning milliy korpusini yaratish uchun quyidagi maqsad va vazifalarni belgilab, o'zimizga reja qilib olishimiz kerak:

1. Mazkur korpus asosiga qanday mantiqiy g'oya qo'yildi?
2. Qanday hajmdagi ma'lumotlar bilan ishlaymiz?
3. Bu matnlar bizga nima maqsadda kerak?
4. Belgilangan matnlarning muhimlik darajasi qanday?
5. Biz matnlarning to'liq shaklidan foydalanamizmi yoki ulardan parchalar biz yaratayotgan korpus uchun yetarlimi?
6. Qanday mavzudagi matnlar bilan ishlaymiz?
7. Korpus yaratishda qanday vazifalarni amalga oshirishimiz kerak?

Ma'lumotlarni yig'ish hamda saralash. Til korpusi yaratishda ularning manbasi sifatida raqamlangan, mualliflari esa raqamlanmagan holdagi matnlardan foydalaniladi. Albatta, bu matnlar uchun birinchi manba allaqachon raqamlanib bo'lingan internet olamidir. Chunki aynan internet korpus uchun titan matnlar

majmuidir. Birinchi navbatda shuni unutmaslik kerakki, bular albatta, veb-sahifada va boshqa internet kanallarida: electron pochta, ICQ va boshqa messenjerlardagi muloqotlar, ijtimoiy to‘rlardagi chatlar va boshqalarda aylanib yuradigan katta hajmdagi matnlardir. Agar tuzilayotgan korpuslar mavzusi va maqsadi qamrab ololsa, electron holdagi boshqa manbalardan ham foydalanish mumkin. Ovozli ma’lumotlarni kompyuter dasturlariga kiritish mumkin bo‘lsa (korpus og‘zaki bo‘lgan taqdirda), mazkur til korpusi juda qiziq bo‘ladi.

Matn hajmi va uni kodlashtirish. Korpus uchun hozirlangan mantlar oddiy tekst formatida bo‘ladi(plain text, *.txt).

Birinchidan, u murakkab MS Word format tipidan ko‘ra kichik joyni egallaydi.

Ikkinchidan, garchi zamonaviy korpus analizi dasturlari odatda, HTML (XML) formatidagi hujjatlar bilan ishlasa-da, har holda bu oddiy tekstdan ko‘ra ishonchliroq. Plain text — bu oddiy ichillikdagi harflar, probel(ochiq joy)lar va tinish belgilari. Har joyda va hamisha, haq qanday dastur tushunadigan fayllar bo‘ladi, zarur bo‘lganda har qanday paytda siz ularni belgilab olishingiz va o‘zingiz tanlagan boshqa biror formatda saqlashingiz mumkin.

Yana bir nozik masala – bu fayllarni kodlashdir. Amallar, jumladan, kompyuterda bajariladigan amallar ingliz tilida ishlash uchun qulay (aniqroq qilib aytganda, lotin alifbosi bilan). Shu sababdan juda ko‘p muammolar boshqa alifbodagi belgilarni kompyuter dasturda tan lmasligi bilan bog‘liq bo‘lmoqda (masalan, rus tili uchun krill alifbosida). Ko‘pgina misollar internet sahifasini belgilaganda ko‘rinadi, unda har qanday tekstni o‘qimasdan, uning o‘rniga ekranga bema’ni tartibdagi ketma-ketlikda krill harflari chiqadi. Bu nafaqat “kodlash” (inglizcha encodings) deb nomlanuvchi voqelikda sodir bo‘ladi, balki ba’zilari rus alifbosida - **koi8-r** yoki **cp1251** orasida ifodalanadi. Ulardan birini qolip sifatida tanlash to‘g‘ri bo‘lmaydi. Bundan tashqari, dunyodagi hamma tillarni va ular foydalanadigan yozuvlarni, hatto misr ierogliflarini ham qo‘llay oladigan Unicode kodlash tizimi paydo bo‘lganiga ko‘p bo‘lgani yo‘q. Yaratilgan barcha programmalar hali u bilan ishlashga tayyor emas.

Har qanday matnli fayl yagona kod bilan saqlanadi. Shunga mos ravishda, agar korpus analiz dasturi yagona kodni hisobga olsa, biz ishlatmoqchi bo‘lgan belgilar jamlanmasini o‘qimasligi mumkin. Bunday muammo faqat ingliz tilidagi matnlarda yo‘q. Ular normal holda o‘qiladi, kodlanadi va analiz qilinadi.

Korpus yorlig‘i (razmetkasi, belgisi, annotatsiyasi). Siz o‘z matningizni belgilashingiz, unga o‘ziga xos belgilar qo‘shishingiz mumkin. Razmetka sizga tekstdagi maxsus joyni aniqlashtirib olishga yordam beradi. Razmetka bu tinish belgilari orqali matnlarni ixchamlashtirishdir.

Korpusni saqlash va uni taqdim etish. Korpus sifatida tugallangan matnlar massivi barcha standartlarga muvofiq bo‘lishi va uning buyurtmachisiga ko‘rsatishga tayyor holda bo‘lishi kerak. Korpusni rasmiylashtirish ham qat’iy qoidalarga ega.[17]

[17] Кутузов А.Б. Корпусная лингвистика. – М., 2005. С. 37

XULOSA

Biz mazkur “O‘zbek tilining milliy korpusini yaratish tamoyillari” mavzusida yozilgan bitiruv malakaviy ishini yozish jarayonida quyidagi xulosalarga keldik:

1. Til korpusi ma’lum tilning belgilangan davrdagi, xilma-xil janr, rang-barang uslub, hududiy hamda ijtimoiy variantdagi matnlarning elektron shaklli maxsus dasturiy ta’minot asosidagi yig‘indisidir. Korpus matnlar massividan iborat bo‘lib, bu matnlar oddiy elektron kutubxonadan farq qiladi. Korpusdagi matnlar maxsus qo‘shimcha ma’lumot bilan boyitilgan va lingvistik tadqiqot uchun asos vazifasini o‘taydi. Tilshunoslikka oid tadqiqotlarda dalil bilan ish ko‘riladigan hollarda o‘sha faktlar yig‘ilishi va sistemaga solinishi lozim.

2. Foydalanishda deyarli kasbiy tabaqalanishga yo‘l qo‘ymaydigan til korpuslari barcha soha vakillarini birday qiziqtirishi tabiiy. Ayniqsa, ona tili va chet tillarni o‘qitish hamda o‘rganish borasida korpuslarning o‘rni beqiyos. Bugun dunyo miqyosida til o‘rgatish tizimi korpuslarga yo‘naltirilayotgani ham fikrimiz dalilidir. Shuning uchun ta`lim, sheva matnlari, poetik matnlar, og‘zaki, ilmiy, rasmiy matnlar korpuslari kabi qator mikrokorpuslar tuzilayotir.

3. Dunyo tilshunosligida qilingan korpus lingvistikasiga oid tadqiqotlar tahsinga sazovor. Ular o‘z tilining milliy korpuslarini yaratish asnosida butun dunyo bilan bog‘lanmoqdalar hamda ushbu tilni o‘rganuvchilar uchun qator yengilliklar yaratilmoqda.

4. Tadqiqot ishini bajarish jarayonida shunga amin bo‘ldikki, korpusni insoniyat hali korpus lingvistikasi fani paydo bo‘lmasdan oldin, ya’ni XVIII asrdan tadqiq eta boshlashgan. Bibliyani tadqiq etish, lug‘atlar yaratish (Johnson, Oxford English Dictionary, Webster Dictionary), tillarni o‘qitish (chastotali lug‘at Thorndike'a, 1921), deskriptiv grammatika (Fries, 1940, Quirk, 1968) va boshqalar bizning mazkur fikrimizni dalillaydi.

5. Tilimizning ichki imkoniyatlari axborot-texnologiyalari tizimida yetarli darajada ishga solinmayotganligi o‘zbek tilining milliy korpusini tuzishni qiyinlashtirmoqda. Yurtimizda o‘zbek tili korpuslarini yaratish ishlarini qoniqarli

deb bo‘lmaydi. Bu ona tilimizning axborot texnologiyalari tili sifatida, shuningdek, milliy korpus shaklida global korpuslar to‘riga ulanib, jahonga chiqishiga monelik qilyapti.

6. Bu sohada olib borilayotgan ilmiy tadqiqot ishlari o‘zbek korpus lingvistikasining nazariy asoslarini ishlab chiqish, namuna sifatida bir-ikki korpus lavhasi yaratish bosqichida turibdi, xolos. O‘zbekistonda amaliy tilshunoslikning bir yo‘nalishi sanalgan o‘zbek korpus lingvistikasini rivojlantirish, uni ilmiy-nazariy jihatdan tadqiq etish, kompyuter va korpus lingvistikasi markazi yoki laboratoriyasini tashkil qilish hamda unga mutaxassislarni biriktirish orqali amalga oshishi mumkin.

FOYDALANILGAN ADABIYOTLAR RO‘YXATI

I. Siyosiy adabiyotlar

1. O‘zbekiston Respublikasi Prezidentining 2003-yil 11-dekabrdagi “Axborotlashtirish to‘g‘risida”gi qonuni.

2. O‘zbekiston Respublikasi Prezidentining 2012-yil 1-fevraldagi “Joylarda kompyuterlashtirish va axborot-kommunikatsiya texnologiyalarini yanada rivojlantirish uchun shart-sharoitlar chora-tadbirlari to‘g‘risida”gi qarori.

3. O‘zbekiston Respublikasi Prezidentining 2002-yil 30-maydagi “Kompyuterlashtirishni yanada rivojlantirish va axborot-kommunikatsiya texnologiyalarni joriy etish to‘g‘risida”gi farmoni.

4. Mirziyoyev Sh.M. Erkin va farovon, demokratik O‘zbekiston davlatini birgalikda barpo etamiz. -Toshkent: O‘zbekiston, 2016.

II. Ilmiy-nazariy adabiyotlar

1. Азарова И.В., Алексеева К.Л., Захарова Л.А. Разметка текстовых фрагментов в корпусе агиографических текстов СКАТ // Труды международной конференции «Корпусная лингвистика – 2006». – СПб: изд-во С.-Петерб. ун-та, Изд-во РХГА, 2006..

2. Андрющенко В. М. Концепция и архитектура Машинного фонда русского языка. М.: Наука, 1989.

3. Беляева Л.Н. Лексикографический потенциал параллельного корпуса текстов // Труды международной конференции «Корпусная лингвистика – 2004». – СПб., 2004.

4. Богданова С.Ю. Исследование слова и предложения компьютерными методами // Слово в предложении: кол. монография / Под ред. Л.М. Ковалевой (отв. ред), С.Ю. Богдановой, Т.И. Семеновой. – Иркутск: ИГЛУ, 2010.

5. Борисова Е.Г. Слово в тексте. Словарь коллокаций (устойчивых словосочетаний) русского языка с англо-русским словарем ключевых слов. – Москва, 1995.

6. Гарабик Р., Захаров В.П. Параллельный русско-словацкий корпус // Труды международной конференции «Корпусная лингвистика – 2006». – СПб., 2006.

7. Гвишиани Н.Б. Практикум по корпусной лингвистике: Учеб. пособие по английскому языку. – М.: Высшая школа, 2008.

8. Герд А.С. РНК и академическая лексикография // Труды международной конференции «Корпусная лингвистика – 2006». – СПб.: Изд-во С.-Петерб. ун-та; Изд-во РХГА, 2006.

9. Герд А.С. Несколько слов о Специальном корпусе текстов (СКТ) // Труды международной конференции «Корпусная лингвистика – 2006». – СПб.: Изд-во С.-Петерб. ун-та; Изд-во РХГА, 2006.

10. Гришина Е.А., Савчук С.О. Национальный корпус русского языка как инструмент для изучения вариативности грамматических норм // Труды международной конференции «Корпусная лингвистика – 2008» 6-10 октября 2008 г. – СПб., 2008.

11. Засорина Л.Н. (ред.). Частотный словарь русского языка. – М., 1977.

12. Захаров В.П. Корпусная лингвистика: Учебно-метод. пособие. – СПб., 2005.

13. Захаров В.П. Веб-пространство как языковой корпус // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог-2005" (Звенигород, 1-6 июня 2005 г.) – М., 2005.

14. Недовшина Е.В. Программа для работы с корпусами текстов: обзор основных корпусных менеджеров. Работа с системой DDC // Языковая инженерия: в поиске смыслов. (электронный ресурс).

15. Rahimov A. Kompyuter lingvistikasi asoslari. – T., 2011.

16. Po‘latov A., Muhamedova S. Kompyuter lingvistikasi. – T., 2007.

17. Xamrayeva Sh. Til korpusining lingvistik va boshqa sohalardagi ahamiyati // O‘zMU xabarlari, – T., 2017.

18. Xamrayeva Sh. Til korpusining leksikografik ahamiyati // O‘zMU, - Toshkent, 2017.

19. U.Tursunov va b. O‘zbek adabiy tili tarixi. Toshkent, “O‘qituvchi”, 1995.

20. McEnery T, Wilson A. Corpus Linguistics. Edinburgh: Edinburgh University Press, 2nd edition, 2001.

21. Mengliyev B va boshqalar. O‘zbek tilining milliy korpusi//Ma’rifat, - Toshkent, 2018.

22. Teubert, W. Corpus Linguistics and Lexicography. International Journal of Corpus Linguistics. Special issue, 2001.

23. Cieri, C., Liberman, M. Language resources creation and distribution at the Linguistic Data Consortium // Proc. of Language Resources and Evaluation Conference (LREC02), 2002.

24. Михайлов М.Н. Контекстносвободная лемматизация как временное решение насущных проблем // Алфавит. Смоленск, СПГУ, 2002.

25. Венцов А.В., Касевич В.Б. Словарь для модели восприятия речи //Вестник С.Петербургского унта. 1998.

III. Internet manbalari

1. Wikipedia.uz
2. www.natlib.uz
3. www.kutubxona.com
4. www.ziyouz.com
5. www.ziyonet.uz
6. www.kh-davron.uz
7. www.ziyouz.com kutubxonasi

8. https://kitobxon.com
9. http://ruscorpora.ru
10. http://www.natcorp.ox.ac.uk/
11. http://sara.natcorp.ox.ac.uk/
12. http://bokrcorpora.narod.ru/
13. http://www.dialog 21.ru
14. http://titania.cobuild.collins.co.uk/
15. http://www.hd.uib.no/icame/brown/bcm.html