# KORPUS VA KORPUS LINGVISTIKASI

**Xidirov Otabek Joʻraboyevich**
Jizzax davlat pedagogika universiteti dotsenti, f.f.f.d (PhD)

## ANNOTATSIYA

Korpus va korpus lingvistikasi haqida nazariy ma’lumotlar keltirilgan hamda oʻzbek tilshunosligida sintaksis boʻyicha yaratilgan nazariyalar asosida korpus uchun parser dasturi ishlab chiqish mumkinligi asoslangan. Sintaksis boʻyicha yaratilgan barcha nazariy materiallarni toʻplash, oʻrganish, umulashtirish hamda sintaktik teglar yaratishda foydalanish zarur.

**Kalit soʻzlar:** Korpus, korpus lingvistikasi, sintaksis, parser dasturi, razmetkalangan, Ixtisoslashgan korpus, ta’limiy korpus, qiyosiy korpus.

Korpus va korpus lingvistikasi XX asrning 60-yillarida paydo boʻldi. Korpus atamasi koʻp ma’noli boʻlib, tilshunoslikda turli ob’ektlarni atab keladi. Umuman olganda, korpus − ma’lum tilga oid materiallar yigʻindisi: u toʻliq matn yoki uning katta fragmentini qamrab oladi. K.Baush korpus atamasiga quyidagicha ta’rif beradi: “Korpus faqat (yozma yoki ogʻzaki) matnlar yoki ularning qismlaridan iborat cheklangan miqdordagi til materialidir”[1]. Tuscan Word Centre xodimi Djon Sinkleyrning ta’kidlashicha, korpus til yoki lingvistik xilma-xillikni tadqiq etish, ma’lumot manbai sifatida namoyish etish uchun koʻrinadigan mezonlarga muvofiq tanlangan elektron koʻrinishdagi matnlar fragmentidan iborat.

Ingliz adabiyotlarida razmetkalangan va razmetkalanmagan korpuslar farqlanadi. Razmetkalanmagan korpus – qayta ishlanmagan, oddiy, “xom” matn, lingvistik informatsiyaga ega boʻlmagan korpus. Masalan, present soʻzi ma’lum kontekstda ot boʻlib kelishini koʻrsatuvchi aniq fakt yoʻq. Bu ma’lumot tushuniladi, ammo yuzaga chiqarilmaydi. Razmetkalangan korpus oʻzida lingvistik axborot saqlaydi: Masalan, present – noun/verb, ya’ni present soʻzi ham ot, ham fe’l boʻlib kela oladi. Lankaster universiteti professori Richard Syao bunday korpusning oddiy matndan farqlanishini atroflicha sharhlab bergan[2]. Oʻz navbatida, Lankaster universiteti olimlari Toni Makenri, Endryu Uilson korpus tilshunosligi (korpus lingvistikasi) toʻgʻrisida quyidagi fikrlarni aytishgan: “Korpus lingvistikasi - bu tilni oʻrganish boʻlib, yozma yoki ogʻzaki oʻqiladigan matnni qayta ishlash, foydalanish va tahlil qilish bilan bogʻliq barcha jarayonlarni oʻz ichiga oladi. Korpus tilshunosligi nisbatan zamonaviy atama boʻlib, “jonli til”dan foydalanishga asoslangan metodologiyaga asoslaniladi. Hozirgi vaqtda korpus tilshunosligining samaradorligi va ahamiyatliligi kompyuter

lingvistikasining rivojlanishi bilan chambarchas bog‘liq. Har qanday matn tarkibini yaratish tamoyillari uning amaliy maqsadiga bevosita bog‘liq[3].

Mutaxassislar izohlanish (teglanish) darajasiga ko‘ra korpusning turli ko‘rinishlarini farqlashadi. Quyida shunday korpus turlariga qisqacha to‘xtalamiz.

1. Ixtisoslashgan korpus − ma’lum bir turdagi matnlar to‘plami: gazeta matni, ilmiy maqolalar.

2. Umumiy korpus turli xil matnlarni o‘z ichiga oladi, matn mazmuni va janriga alohida talab o‘yilmaydi.

3. Qiyosiy korpus. Ular turli tillarning ikki yoki undan ko‘p kichik qismlari, masalan, rus hamda nemis tili yoxud bitta tilning variantidan iborat. Masalan, Avstriya va Shveysariya nemis tili versiyalari.

4. Parallel korpus turli tillardagi o‘xshash matnlarni o‘z ichiga olgan ichki korpusdan iborat. Birinchi korpusdan asliyatdagi matn, ikkinchisidan tarjima matn o‘rin oladi.

5. Ta’limiy korpus − bu chet tilini o‘rganayotgan shaxs uchun ona tilida so‘zlashuvchilar tomonidan yozilgan matnlar to‘plami.

6. Didaktik korpus chet tilini o‘qitish jarayonida foydalaniladigan til ma’lumotlaridan iborat.

A.N.Baranov korpus texnologiyalariga quyidagicha ta’rif beradi: “Korpus − kompyuter tilshunosligining eng muhim vositalaridan biri. Ular tilni tahlil qilish, matnni lingvistik korpus shaklida taqdim etishning amaliy vazifalarini hal qilishga imkon beradi”[4].

M.K.Maxmutovaning fikricha, dastlab, tillarni tadqiq qilishda korpusdan foydalanishning maqsadi turli til elementlarining chastotasini hisoblab chiqishdan iborat edi. Bunday elementlar so‘z, so‘zshakl, morfema va iboralar bo‘lishi mumkin. Korpusdan til va nutq birliklari bo‘yicha turli xil ma’lumot va statistika olish uchun foydalanish mumkin[5]. Bunday texnologiya, xususan, leksikografiya, so‘zni avtomatik qayta ishlash tizimlari sohasida turli vazifalarni hal qilish imkonini beradi. Nutqni aniqlash, sintez qilish, avtomatlashtirilgan va mashina tarjimasi, imlo va grammatikani tekshirish kabi murakkab lingvistik muammolarni hal qilishda statistik usuldan ham foydalaniladi. Masalan, korpus materialida qaysi so‘zning turg‘un iboralar guruhiga tegishli ekanligini aniqlash mumkin. Buning uchun olingan ma’lumotda, birliklar o‘zaro muntazam ravishda birika olishini tekshirish kerak.

Gollandiyadagi Nijmegen universitetida ishlab chiqilayotgan grammatikalar korpus matni holatlarida sinovdan o‘tkaziladi. Grammatika asosida korpusni qayta ishlaydigan tahlil dasturi tuziladi. Olingan ishlov berish natijalari grammatika ma’lumotlarini qanchalik aniq tasvirlashini ko‘rsatadi. Shundan kelib chiqqan holda,

korpus texnologiyasi tilshunoslikning yangi nazariyalari, so‘zni avtomatik qayta ishlash tizimlarini sinab ko‘rishga imkon berishini ko‘ramiz.

1993 yilda Lancaster-Oslo/Bergen (LOB) korpusi va Britaniya milliy korpusi (BNC) yaratuvchisi Jeffri Leich tomonidan 1993 yilda tuzilgan ANNOTATSIYAlash postulatlardan biri til belgilarining aniq va tushunarli tavsiflash prinsipi e’tiborga molik. Shuningdek, uning fikriga ko‘ra, umumfoydalanishga mo‘ljallangan korpusning razmetkasi uchta prinsipga muvofiq bo‘lishi kerak.

1. Razmetka foydalanuvchi uchun qo‘llanma yoki ko‘rsatma shaklida mavjud bo‘lgan tahlil sxemasiga asoslangan bo‘lishi, har bir parametr undan joy olishi kerak.

2. Foydalanuvchi uchun ochiq korpus razmetkasi “nazariy jihatdan neytral” bo‘lishi lozim: razmetka parametrlari barcha uchun tushunarli bo‘lgan tushunchalar tizimidan tashkil topgan bo‘lishi talab etiladi. Agar korpus aniq bir loyiha uchun mo‘ljallangan bo‘lsa, uni razmetkalashda maxsus, aynan muallifga xos, umumqabul qilingan tasnifdan foydalanish lozim: bunda ham tuzuvchidan u yoki bu til nazariyasiga tayanish talab qilinadi.

3. Korpus annotatsiyasi sxemasi kim tomonidan, qaysi auditoriyaga mo‘ljallanganligi aniq, ravshan ko‘rsatilishi lozim, chunki korpusdan foydalanishda yuridik va texnik jihatdan turli chegaralar mavjud[6].

J.Leichning birinchi postulati mukammal ishlangan, ammo u hamma korpusda ham yuzaga chiqavermaydi. Albatta, barcha korpuslar u yoki bu darajada ma’lumot(teg)lar tizimi bilan ta’minlangan bo‘ladi. Har bir parametrning qanday ma’lumot tashiyotganligini aniqlashga doim ham erishib bo‘lmaydi. Bu borada “matnni avtomatik qayta ishlash” (http://www.aot.ru) guruhida faoliyat olib borgan olimlarning faoliyatini alohida ta’kidlash lozim. Ular asosan, rus tilidagi matnlarga avtomatik ishlov berish jarayonlarini ishlab chiqishgan; kompyuter texnologiyalari yutuqlari nazariy tilshunoslik bilan birlashtirib, amaliy natijalarga erishilgan.

Demak, har qanday sintaktik teglar tizimini ishlab chiqish uchun kompyuter texnologiyalari yutuqlari bilan birgalikda o‘zbek tilshunosligida sintaksis bo‘yicha yaratilgan nazariyalar asosida korpus uchun parser dasturi ishlab chiqish mumkin. Sintaksis bo‘yicha yaratilgan barcha nazariy materiallarni to‘plash, o‘rganish, umulashtirish hamda sintaktik teglar yaratishda foydalanish zarur.

## FOYDALANILGAN ADABIYOTLAR RO‘YXATI: (REFERENCES)

**1.** Maxmutova M.K. Problemы annotirovaniya (tagirovaniya) tekstov v korpusnoy lingvistike // Vыpusknaya kvalifikatsionnaya rabota YUUrGU. – Chelyabinsk, 2018. – 94 s.

**2.** Xianyao Hu, Richard Xiao. How do English translations differ from non-translated English writings? A multi-feature statistical model for linguistic variation analysis // https://www.researchgate.net/publication/294138492_How_do_English_translations_differ_from_non-translated_English_writings_A_multi-feature_statistical_model_for_linguistic_variation_analysis/references

**3**. Maxmutova M.K. Problemы annotirovaniya (tagirovaniya) tekstov v korpusnoy lingvistike // Vыpusknaya kvalifikatsionnaya rabota YUUrGU. – Chelyabinsk, 2018. – 94 s.

**4.** Baranov, A. N. Vvedenie v prikladnuyu lingvistiku [Tekst]: uchebnoe posobie / A.N. Baranov. – M.: Izd-vo Editorial URSS, 2001. – 347 s.

**5.** Chernyakova, T. A. Metodika formirovaniya navыkov studentov na osnove lingvisticheskogo korpusa [Tekst] / T. A. Chernyakova. – Tambov, 2012. – 149 s.

**6.** Leech, G. Corpus ABSTRACT schemes / G. Leech Literary and Linguistic Computing, 1993. – 8/4. – P. 275-281.