Korpus – leksikografik ma’lumotlar bazasi uchun lingvistik instrument sifatida

O‘zMU mustaqil izlanuvchisi
O.O Yulbarsov

Annotatsiya. Maqolada o‘zbek tilshunosligida nisbatan yangi soha bo‘lmish Kompyuter lingvistikasining ahamiyati hamda uning amaliyotga tatbiq qilish zarurati haqida qarashlar ilgari surilgan. Xususan, Korpus – leksikografik ma’lumotlar ba’zasi uchun muhim lingvistik instrument ekanligi ilmiy asoslantirilgan. Kompyuter lingvistikasining ajralmas tarmog‘i sifatida Korpus lingvistikasi alohida – konkret nazariyalarga ega soha. Maqola orqali ilmiy-nazariy qarashlar, zamonaviy sohalardagi o‘zgarishlarni amaliyotga tatbiq etishdagi muammolar sanab o‘tilgan.

Kalit so‘zlar: Korpus, korpus lingvistikasi, leksikografik ma’lumotlar ba’zasi, lingvistik instrument, NLP, Kadling, Dougles Biber, A. Po‘latov, N. Yoqutova, M. Ayimbetov, S. Rizayev, www.uzbekcorpus.uz, lug‘at.

Zamonaviy tilshunoslikning hozirgi kundagi global vazifalaridan biri – tilni tadqiq qilish va uni amaliyotga tatbiq etishda kompyuter texnologiyalaridan maqsadli foydalanishdan iborat. Tilshunoslikka kirib kelayotgan har bir yo‘nalish yoki terminlarni o‘z sohasi o‘laroq tasnif etish, undan amaliy foydalanish, o‘zbek tilini axborot-kommunikatsiya texnologiyalari orqali qo‘llash kabi ishlar til tadqiqining ustuvor maqsadlari safiga qo‘shishni taqozo etadi. Xususan, Kompyuter lingvistikasining ajralmas qismi hisoblanadigan Korpus lingvistikasi sohasi ham tilning amaliy – matn aspektidagi muhim instrumentlardandir. Korpusni lingvistik instrument sifatida qarash hozirgi kundagi zamonaviy lingvistika sohasining ijtimoiy tilga moslashuvchanligiga misol bo‘ladi deyish mumkin, bizningcha.

XXI asrning global muammolaridan biri tabiiy tillarning milliy xususiyatini saqlab qolishdan iborat. Dunyo tillarini elektron korpuslarini yaratish va rivojlantirishda NLP hamda til texnologiyalariga doir tadqiqotlarni izchil ravishda olib borish dolzarb vazifaga aylandi.[1] Demak, bugungi kunda Korpus tushunchasi o‘z funksiyalarini til va Tilshunoslik sohasida amaliy tarzda taqdim etuvchi sifatida namoyon bo‘lmoqda. Bu terminni tilimizda juda ko‘p o‘rinlarda ishlatib kelmoqdamiz.

Wikipediya ma’lumotiga ko‘a Kopus (lotincha: corpus – tana) – 1) inshoot, mexanizm, asboblar, apparat, qurilmalar va ba’zi transport mashinalarining qismi; 2) yaxlit maydondagi bir necha binodan bittasi yoki yirik inshootning alohida qismi;

[1] Abduraxmonova N. O‘zbek tili elektron korpusining kompyuter modellari (monografiya). / Toshkent:Muharrir, 2021, 202 b.

3) kegeli (o‘lchami) 10 punktga (3,76 mm ga) teng bo‘lgan bosma shrift; 4) biror maxsus ish bilan mashg‘ul bo‘lgan mutaxassislar hay’ati (mas, diplomatik korpus, ofitserlar ko‘rpusi va boshqalar).[2]

Korpus – bu matnlar majmuasi. Ya’ni ma’lum sohadagi matnni lingvistik nuqtayi nazardan tasniflovchi elektron to‘plam. Bu mantlar turli jihatga ko‘ra tasniflangan holatda bo‘lishi ham mumkin. Ular kompyuterning ma’lumotlar bazasida saqlanadi. Matnlar og‘zaki va yozma shaklda bo‘ladi.

Ilk korpusning yaratilishi 1812-yilga borib taqaladi, bunda nemis olimi Kadling o‘zining nemis so‘zlaridagi undosh harflar distributsiyasini tahlil qilgan. Vaholangki, u davrda hali kompyuter terminining o‘zi ham bo‘lmagan. Keyinchalik zamonaviy ingliz tili korpusi namunalaridan biri sifatida Broun korpusi 1960-1961 yillarda yaratildi va u ilk bor bosma holda chop etildi. Oradan bir yil o‘tibgina ushbu korpus elektronlashtirildi. Ko‘rinib turibdiki, ilk korpuslar kompyuter texnologiyalarisiz ham mavjud bo‘lgan va izlanishlar olib borilgan. Keyinchalik fan rivoji qidiruv metodi (konkordans)ni elektron formatdagi matnlarda o‘tkazishni taqazo etdi va kompyuter lingvistikasi bilan sohalararo munosabatga ehtiyoj tug‘ildi.

Shu bilan birga kompyuter tilshunosligi odatda, kompyuter vositalari (dasturlari, ma’lumotlarni tashkil qilish va qayta ishlash uchun kompyuter texnologiyalari)ni muayyan sharoitlarda, vaziyatlarda, muammoli sohalarda va tillardagi modellarning ko‘lamini nafaqat tilshunoslikda, balki boshqa fanlarga ham qo‘llashni nazarda tutadi.[3][№2-2019/. 38 б]

Korpus tushunchasi hozirgi kundagi lingvistik tadqiqotlarda, xususan, amaliy tilshunoslikda faol qo‘llana boshlaganini uchratishimiz mumkin. Ko‘rishimiz mumkinki, bu atamaning tilda fan tarmog‘iga aylanib borishi aynan tilning material sifatidagi yo‘nalishlari uchun xususiy tahlillarga olib kelgan. Shuning uchun ham bugungi kunda Kompyuter lingvistikasining ajralmas tarmog‘i sifatida Korpus lingvistikasi alohida – konkret nazariyalarga ega soha deyish mumkin. Ko‘rpus lingvistikasini bir qancha funksiyalari mavjud bo‘lib, ulardan eng muhimi matnli resurslarning kompyuter bazasini yaratishdan iborat. Matnlar bazasi esa har xil sohalar uchun ma’lumotlarni tezkor qidirib topish hamda undan o‘z tadqiqotlarida foydalanishda olimlarga juda katta yordam bermoqda.

Korpus lingvistikasining hozirda alohida soha yoki fan bo‘lib shakllanishida olimlarning nazariy qarashlari katta hissa qo‘shgan. Korpus lingvistikasi asosan Kompyuter lingvistikasi bilan birgalikda uning bir yo‘nalishi deb baholangan bo‘lsa, ayrim nazariy qarashlarda alohida funksiyaga ega fan sifatida o‘rganiladi. Shuningdek Korpus lingvistikasi sohasida nazariy hamda amaliy tadqiqotlar olib

---

[2] https://uz.wikipedia.org/wiki/Korpus

[3] Н.Б. Атабоэв. Корпус лингвистикасининг асосий хусусиятлари. www.journal.fledu.uz/ “Ўзбекистонда хорижий тиллар” илмий-методик электрон журнал/ №2-2019/. 38 б.

borgan olimlar soha rivojiga salmoqli hissalarini qo‘shganlar. Xususan, Kading, Fries va Traver, Bongers, Kennedy, Stefan Gries, L.N. Zasorin, V. Zaxarov, Dougles Biber kabi dunyo olimlari hamda O‘zbekistonda Kompyuter lingvistikasi sohasi shakllanishiga ulkan hissa qo‘shgan professor A. Po‘latov, ilmiy tadqiqotlar olib borgan N. Yoqutova, M. Ayimbetov, S. Rizayev va S. Muhamedov va boshqalar ilmiy-amaliy ishlar olib borgan.[4] Bundan tashqari N.Z.Abduraxmonova “Kompyuter lingvistikasi” nomli darsligida tabiiy tillarni qayta ishlashning lingvistik asoslarini yaratish hamda korpus lingvistikasiga oid nazariy va amaliy masalalarni o‘zbek tili doirasida o‘rganadi. Olimaning muallifligida uzbekcorpus.uz – o‘zbek tilining elektron korpusi yaratildi. Bu korpusni ayni paytdagi holatida o‘zbek tilida yaratilgan lug‘atlar, web sahifalar, o‘zbek tilining morfologik ma‘lumotlar bazasi, o‘quv adabiyotlar hamda turli janrdagi ilmiy, rasmiy va badiiy matnlar majmuasi tashkil topgan.[5] Nazariy hamda amaliy tadqiqotlar o‘zbek tili korpusi uchun muhim qirralarini jadallik bilan kashf qilib bormoqda.

Korpus lingvistikasi haqida olimlarning nuqtayi nazarlari turlicha. Qayd etilishicha, u til modeli emas, muayyan darajada metodologik yondashuv sifatida qarash to‘g‘ri bo‘ladi. Dougles Biber uni quyidagi sifatlarini sanab o‘tadi:

- u tabiiy matndagi zarur birliklarni empirik tahlil qiladi;
- analiz uchun “Korpus” sifatida tabiiy matnlarning katta va tizimlashtirilgan jamlanmalarini birlashtiradi;
- analiz uchun kompyuterning ham avtomatik va interaktiv texnologiyalardan foydalanish imkonini beradi;
- u analitik texnologiyaning miqdor (statistik) va sifat xususiyatlarini o‘z ichiga oladi.[6]

Aytib o‘tganimizdek Korpus tilning har qaysi sohadagi turli maqsadlari uchun faol xizmat qiladigan zamonaviy lingvistik texnologiya. Tilshunoslikda ham korpusni tilning funksional jihatdan tasniflovchi muhim qurilma deyishimiz mumkin. Ayniqsa lug‘atlarda hamda leksikografik ma’lumotlarni shakllantirishda samarali instrument deb qarashimiz mumkin.

Yuqorida sanab o‘tilgan olimlarning qarashlarini hamda Korpus lingvistikasining tilda dolzarblik tendensiyalarini o‘rganar ekanmiz, bu omillar lingvistikaning boshqa sof elementlarini ham elektron korpusga kiritish, tahlil qilish zaruratini yuzaga keltirmoqda. Xususan, tilshunoslikdagi leksikografiya tushunchasi va uning amaliy ko‘rinishi ham zamonaviy amaliy kompyuter leksikografiyasining rivojlanishi uchun muhim manba bo‘la oladi. Lug‘atning har qanday ko‘rinishi hozirda elektron tarzga ko‘chib borayotganligi undan foydalanishda qulayliklarni

[4] N. Z. Abduraxmonova. “Kompyuter lingvistikasi” Darslik. Toshkent – 2021/ 4-, 15-, 284-, 285-b.
[5] www.uzbekcorpus.uz
[6] N. Z. Abduraxmonova. “Kompyuter lingvistikasi” Darslik. Toshkent – 2021/ 284 b.

yaratib bermoqda. Korpusdagi leksikografik ma’lumotlar bazasi nima ekanligini tushunishdan avval leksikografiya terminini tilshunoslik nuqtayi nazariga ko‘ra bilishimiz kerakligini taqazo etadi.

Leksikografiya yunoncha “leksiko” va “grafiya” so‘zlaridan tashkil topgan bo‘lib, tilshunoslikning lug‘atchilik bilan shug‘ullanuvchi sohasidir. Lug‘atlar tildagi so‘zlarni, ibora va maqollar, turli nomlarni ma’lum tartibda o‘z ichiga olgan majmualar. Bunday lug‘atlar o‘tmishda qo‘lyozma shaklida ham bo‘lgan. O‘zbek leksikografiyasi tarixi Mahmud Koshg‘ariyning “Devonu lug‘atit turk” asaridan boshlangan.

Leksikografiyaning vazifa doirasiga quyidagilar kiradi:

1. lug‘at tuzish prinsiplari va metodikasini ishlab chiqish;
2. lug‘at tiplari va turlarini aniqlash;
3. lug‘atshunoslarning ishini tashkil qilish;
4. lug‘at tuzish uchun asos bo‘ladigan kartoteka fondini yaratish;
5. lug‘atchilik tarixini o‘rganish;
6. lug‘at tuzish bilan shug‘ullanish.[7]

Demak, leksikografiya tilimizda mavjud tushunchalarning ma’no-mazmunini kengroq ifodalab berishga xizmat qilar ekan. Dastlabki yaratilgan lug‘atlarni internet tarmog‘iga yuklash lug‘atshunoslikning yana bir jihati hisoblanadi. Bu esa Korpus lingvistikasining bir funksiyasini namoyon etadi. Korpusga joylangan lug‘atlar ham internet tarmog‘iga joylashtirishda, ham u orqali boshqa lug‘atlarni tezkor tuzishda yordam beradi. Elektron lug‘atlarning tuzishda, ularning joylashuvi va so‘rovlariga ko‘ra ma’lumot berish imkoniyati dastlabki ishlab chiqish jarayonini talab qiladi. Shuning uchun kitob shaklidagi lug‘atlarni elektron tarzda joylashtirishda kompyuterning, xususan, korpusning amaliy-funksional imkoniyatlari hisobga olinadi.

Leksikografik ma’lumotlar bazasi – katta ko‘lamli ma’lumotlar majmui. Ma’lumotlardan foydalanishda tezkorlik va samaradorlik bugungi kun talabidir. Shu bois ma’lumotlardan istalgan vaqtda tezkor foydalanishda korpus muhim instrument bo‘la oladi.

Hozirda tilshunos olimlar tomonidan leksikografiyada kompyuterlardan keng foydalanish turlicha izohlanadi. Ba’zilar bu yerda kompyuterning bir vosita ekanligini, kompyuterning vosita bo‘lib xizmat qilishi leksikografiya uchun muhim emasligini ta’kidlaydilar. Lug‘atlar tuzishda kompyuterning rolini ortiqcha bo‘rttirib

[7] H. Jamolxonov. Hozirgi o‘zbek tili. Toshkent – 2015/

ko‘rsatib, ular sohada inqilobdir, degan fikrni qo‘llab-quvvatlovchilar soni ko‘pchilikni tashkil qiladi.

Maxsus qidiruv tizimiga asoslangan mobil lug‘at korpusi, qidiruv parametrlari har xil turdagi lug‘atlarda mavjud bo‘lgan leksik birlikning lug‘aviy ma’nosi va qidirilayotgan leksema yoki so‘zga tegishli misollarni ma’lumot bazasidan o‘qiy olish imkoniga ega bo‘lishi kerak. Bu borada avvalo korpusga asoslangan lug‘at qanday ishlashini aniqlab olish zarur.

Rus tilshunosi I.I.Sajenin Korpusning xarakterli belgilari sifatida quyidagilarni ko‘rsatib o‘tdi:

- tilga oid ma’lumotlarining yirik massivi;
- elektron ma’lumotlar bazasi;
- birlashtirilishi;
- tuzilishi;
- belgilanganlik;
- filologik kompetentlik;
- maxsus qidiruv tizimining mavjudligi.

Olim ushbu omillar orqali korpusga asoslangan lug‘atlarning ma’lumotlar bazasi hamda dasturiy ta’minotlarni shakllantirishga alohida e’tibor qaratib, til o‘rganuvchilar qiduruv tizimidan foydalangan holda kerakli so‘zning umumiy ma’lumotlarini olish mumkin ekanligini ta’kidlaydi[8].

Korpus boshqa til vakillarini ma’lum tilni yaxshiroq tushunish, nazariy hamda amaliy aniq xulosaga kelishi uchun ham muhim instrument sifatida namoyon bo‘ladi. O‘zbek tilida ham biror so‘zni konteks bilan namoyon bo‘lishi yoki matn tarkibida kelishi til o‘rganuvchilar uchun qulaylik keltirib chiqaradi. Ayni bu amaliyotlarning hammasi Korpus leksikografiyasi sohasi, xususan, korpusda mujassam bo‘ladi. Leksikografik ma’lumotlar bazasi aynan korpus lingvistikasini ajralmas hamda to‘ldiruvchi qismi ekanligi anglashiladi. Korpus lingvistikasining nazariy xususiyatlarini hisobga olgan holda mazmunli lug‘at materiallarini yaratish, leksikografik manbalarni elektronlashtirilish korpusning vazifasi hamda bu borada erishilgan amaliy natijalar zamonaviy korpusning yutuqlaridan deyish mumkin.

[8] И. И. Саженин словарный корпус как элемент оптимизации исследовательского процесса

Foydalanilgan adabiyotlar ro‘yxati

1. Mengliev, D. B., Abdurakhmonova, N., Hayitbayeva, D., & Barakhnin, V. B. (2023, November). Automating the transition from dialectal to literary forms in Uzbek language texts: an algorithmic perspective. In 2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE) (pp. 1440-1443). IEEE.
2. Abdurakhmonova, N., & Urdishev, K. (2019). Corpus based teaching Uzbek as a foreign language. *Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL)*, *6*(1-2019), 131-7.
3. Ismailov, A. S., Shamsiyeva, G., & Abdurakhmonova, N. (2021). Statistical machine translation proposal for Uzbek to English. *Science and Education*, *2*(12), 212-219.
4. Abduvahobov, G. I. (2024). CONCEPTUAL FOUNDATIONS OF ELECTRONIC EDUCATIONAL DICTIONARY. *International journal of artificial intelligence*, *4*(04), 80-82.
5. Abduvahobov, G. (2021). About the concept of computer lexicography. *ISJ Theoretical & Applied Science*, *6*(98), 664-668.