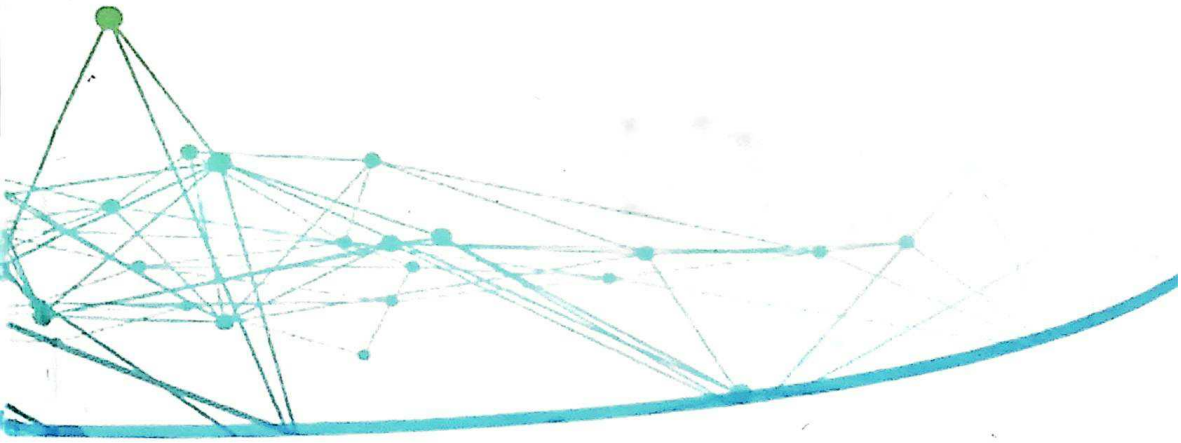
Globe EDIT



Nilufar Abduraxmonova

KORPUS LINGVISTIKASI

darslik

O‘ZBEKISTON RESPUBLIKASI
OLIY TA’LIM, FAN VA INNOVATSIYALAR VAZIRLIGI
MIRZO ULUG‘BEK NOMIDAGI O‘ZBEKISTON MILLIY UNIVERSITETI

ABDURAXMONOVA NILUFAR ZAYNOBIDDIN QIZI

KORPUS LINGVISTIKASI

(Darslik)

70230801 – Kompyuter lingvistikasi magistratura mutaxassisligi

Toshkent-2024

Ushbu darslik 70230801 – Kompyuter lingvistikasi magistratura mutaxassisligidagi magistrlarga mo‘ljallangan bo‘lib, o‘zbek tili elektron korpusini konseptologik va strukturaviy loyihalashda xorijiy tajriba amaliyotini o‘rganish, tilning lingvistik korpusini yaratishda morfologik va sintaktik teglash va tahlil qilishning FST va UdPipe kabi avtomatik usullarini o‘zbek tiliga tatbiq qilish orqali lingvistik algoritmni tuzish hamda lisoniy modellarini mashina tiliga o‘tkazish, matn fragmentining reprezentativligi va qidiriv birliklari (lemma va token)ni tahlil qilish uchun matn korpusining lingvistik va dasturiy ta’minotini tuzish, o‘zbek tili uchun korpus yaratish texnologiyalari va metodlarini lingvistik instrumentariylar yordamida amalga oshirish, korpus menejerining formal-funksional modellari asosida korpus interfeysini shakllantirishga oid bilim va ko‘nikmalarni nazariy va amaliy jihatdan shakllantirishga yordam beradi.

Mas’ul muharrir: f.f.d., prof. H.Dadaboyev

Taqrizchilar:

Mersaid Aripov – fizika-matematika fanlari doktori, Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universiteti professori

Saodat Muhamedova – filologiya fanlari doktori, Alisher Navoiy nomidagi Toshkent davlat O‘zbek tili va adabiyoti universiteti professori

Darslik Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universiteti Kengashining 2023-yil 31-martdagi 8-sonli qaroriga asosan nashrga tavsiya qilingan.

MUNDARIJA

KIRISH

Jahon amaliy tilshunosligida korpusshunoslik, korpus texnologiyasi va korpus lingvistikasi muammolarini o‘rganish XX asrning ikkinchi yarmida rivojlanish bosqichiga ko‘tarildi. XX asrda dastlabki ingliz tili korpusining paydo bo‘lishi boshqa dunyo tillarida ham katta hajmdagi elektron korpuslarning yaratilishida muhim rol o‘ynadi. Endilikda tabiiy tilning lisoniy modellari va nutqiy imkoniyatlarini kompyuter tiliga o‘tkazish, til bilan bog‘liq masalalarni axborot texnologiyalari hamda metodlari yordamida tabiiy tilni o‘rganishda korpus obyekt vazifasini bajarmoqda. Tabiiy tillarning rivojlanish barqarorligi, ularning milliy sofligini saqlab qolishga qaratilayotgan bir davrda tillarning elektron korpuslarini yanada takomillashtirish va yangi texnologiyalarni yaratish axborot asrining dolzarb masalalardan biri hisoblanadi.

Dunyo kompyuter lingvistikasi integrallashgan soha sifatida shiddat bilan rivojlanib borayotgan davrda qator usullar va metodlar orqali til muammolarini hal qilishda muhim vosita bo‘lib xizmat qilmoqda. Kompyuter texnologiyalarining til rivojiga va aksincha, til texnologiyasining kompyuter texnologiyalariga ijobiy ta’siri orqali qator ilmiy yutuqlarga erishildi. Jahonda kompyuter lingvistikasining turli yo‘nalishlari bo‘yicha ko‘plab olimlar tomonidan ilmiy izlanishlar olib borildi[1]. Buning natijasida fanning mashina tarjimasi, korpus lingvistikasi, kompyuter leksikografiyasi, tahrirlovchi dastur, nutqiy sintezator, lingvostatistik tahlil qiluvchi dastur, matnlarni referatlash va tasniflash kabi qator yo‘nalishlari vujudga keldi.

Bugungi kunda o‘zbek kompyuter lingvistikasi, tabiiy tilni qayta ishlash (NLP), mashinali ta’lim (machine learning), ma’lumot qidiruvi (Data Mining)

[1] Jurafskiy D., Martin J. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2007. – P. 12-13; Karin A., Beng A. Advances corpus-based contrastive linguistics. USA: John Benjamins, 2013 – P. 25-54; Koehn P., Och F.J., Marcu D. Statistical phrase based translation // Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL). Proceedings of the Joint Conference 2003; Mitkov R. The Oxford handbook of Computational linguistics. Oxford university press, 2003; Kurdi M.Z. Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax. – Great Britain: Wiley-ISTE, 2016. – 300 p.; Чардин И.С. Лингвистические корпусы с разметкой на основе грамматики зависимостей и их применение при автоматическом синтактическом анализе: Автореф.дисс. ...д-ра филол. наук. – Москва, 2004. – 24 с.

kabi yangi sohalar kesimida rivojlanmoqda hamda ushbu sohalarda erishilayotgan ilmiy-amaliy natijalar bir-birini to'ldirib bormoqda. O'zbekiston Respublikasi Prezidentining 2019-yil 21-oktabrdagi "O'zbek tilining davlat tili sifatidagi nufuzi va mavqeyini tubdan oshirish chora-tadbirlari to'g'risida"gi PF-5850-son Farmoniga muvofiq davlat tilining zamonaviy axborot texnologiyalari va kommunikatsiyalariga integratsiyallashuvini ta'minlashda qator ustuvor vazifalar ko'rsatib o'tildi[2]. O'zbek tilining mavqeini mustahkamlash va uni til siyosati darajasida nufuzini oshirishda tilning raqamli resursini yaratishga doir qator dasturlar ishlab chiqilmoqda.

Dunyo tajribasida korpus yaratishning lingvistik, matematik va dasturiy tomonlari olimlarining tadqiqotlarida o'z ifodasini topgan[3]. Chunonchi, rus va ingliz tillari bo'yicha korpus lingvistikasi turli sohalar kesimida V.Zaxarov, A.Sedov, A.Baranov, R.Potapova, V.Rikov, U.Frensis, N.Leontyeva, V.Martin, S.Kubler, A.Laurens, E.Etwell, S.Hunston, L.Boizou, McKenneri, J.Grafmiller, J.Grieve, N.Grum, S. Hansson, K.McAuliff, M.Malberg, P.Milin, A.Murakami, R.Peych, A.Schembri, P.Tompson, B.Vinter, G.Lich kabi xorijiy olimlar[4] tomonidan hamda turkologiyada turk tili korpusi bo'yicha Aksan, Deniz Zeyrek, Kemal Oflazer, Umut Özge Bular; uyg'ur tili bo'yicha Yusup Aibaidulla, Kim-Teng Lua; boshqird tili bo'yicha L.A.Buskunbaeva, Z.Sirazitdinov; hakas tili

---

[2] O'zbekiston Respublikasi Prezidenti Shavkat Mirziyoyevning 2020-yil 20-oktabrdagi «Mamlakatimizda o'zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to'g'risida»gi PF-6084-son farmoni // https://lex.uz/docs/5058351.

[3] Kubler S., Zinsmeister H. Corpus linguistics and linguistically annotated corpora. – New York: Bloomsbury, 2015. –P. 321.; Martin W. Developing Linguistic Corpora: A Guide to Good Practice, OxfordBooks. 2005. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/; Heike Z., Hinrichs E., Kübler S., Witt A. Linguistically annotated corpora: Quality assurance, reusability and sustainability / Corpus Linguistics: An International Handbook A. Lüdeling and M. Kytö (eds), Vol. 1, – Berlin: Mouton de Gruyter, –P. 759–76; Копотев М. Введение в корпусную лингвистику (учебное пособие). – Прага, 2014. – 264 с.; Atkins B., Zampolli A. Computational approach to the lexicon. – Oxford, 1994. –P. 494.

[4] Седов А.В Математические модели, методы и алгоритмы построения размеченных корпусов текстов: Автореф. дис... канд. тех. наук. – Петрозаводск, 2013. – 22 с.; Anthony L. AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom, 2005. // IEEE International Professional Communication Conference Proceedings, – P.729-737; Atwell E. Development of tagsets for part-of-speech tagging / An international handbook. Corpus Linguistics: Mouton de Gruyter. 2008; Баранов А.Н., Михайлов М.Н., Сидоров Г.О. Динамический корпус текстов как новая технология прикладной лингвистики // Труды международного семинара Диалог'98 по компьютерной лингвистике и ее приложениям. – Т. 1998; Hunston S. Corpora in Applied Linguistics. Cambridge University Press, 2002. – 234 p.

bo‘yicha Sheymovich, tatar tili bo‘yicha J.Suleymanov, A.Gatiatullin, O.Nevzorova, R.Gilmullin, B.Hakimov; qirimtatar tili bo‘yicha L.Kubedinova hamda tuva tili bo‘yicha A.Salchak kabi olimlarning ishlari diqqatga sazovor.

O‘zbekistonda kompyuter lingvistikasining shakllanishida dastlab lingvostatistikaga oid N.Yoqubova, M.Ayimbetov, S.Rizayev, S.Muhamedov kabi olimlar tomonidan izlanishlar olib borilgan[5]. S.Muhamedovning P.P.Piotrovskiy bilan hammualliflikda yozgan «Инженерная лингвистика и опыт системно – статистического исследования узбекских текстов» nomli kitobida lingvistik modellar, modellashtirish va uning tamoyillari, o‘zbekcha matnlarning kvantativ tahlillari o‘rganilgan. Shuningdek, so‘nggi o‘n yillikda kompyuter lingvistikasi sohasida A.Po‘latov, S.Muhamedova, A.Rahimov, Z.Xolmanova, N.Abduraxmonova kabi olimlarning sohaga doir o‘quv darslik va qo‘llanmalari nashrdan chiqdi[6]. Tilni kompyuter modellari, lingvistik ta’minotini yaratish, shuningdek, kompyuter metodlari yordamida lingvistik masalalarini yechishga yo‘naltirilgan monografik tadqiqotlar olib borildi[7].

Korpus lingvistikasida erishilgan natijalar kompyuter lingvistikasi va tabiiy tilni qayta ishlash sohalari uchun ham til texnologiyasi borasida tub burilishlar yasadi. Bunda tabiiy tilning mashina tilini yaratish, til bo‘yicha statistik ma’lumotga ega bo‘lish, sun’iy intellektning lisoniy va nutq modellarini

---

[5] Ризаев С. Ўзбек тилининг лингвостатистик тадқиқи: Филол. фан. д-ри. ... дисс. автореф. – Тошкент, 2008. – 50 б.; Мухамедов С.А. Статистический анализ лексико-морфологической структуры узбекских газетных текстов: Автореф. дисс. ...канд. филол. наук. – Ташкент, 1980. –25 с.; Бектаев К.Б., Пиотровский Р.Г. Математическая лингвистика. – М.: Высшая школа, 1997. – 420 с.; Айымбетов М.К. Проблемы и методы квантитативно-типологического измерения близости тюркских языков (на материалах каракалпакского, казахского и узбекского языков): Автореф. дисс. ...д-ра филол. наук. – Ташкент, 1997. – 47 с.

[6] Пўлатов А. Компьютер лингвистикаси. – Т.: Akademnashr, 2011. – 175 б.; Норов А. Компьютер лингвистикаси асослари. – Қарши, 2017. – 136 б.; Муҳаммедова С. Ҳаракат феъллари асосида компьютер дастурлари учун лингвистик таъмин яратиш. Методик қўлланма. – Тошкент, 2006.; Холманова З. Компьютер лингвистикаси (ўқув қўлланма) –Тошкент, 2019.; Abduraxmonova N.Z. Kompyuter lingvistikasi (darslik). – Toshkent: Nodirabegim, 2021. –398 b.

[7] Абдурахмонова Н.З. Инглизча матиларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти (содда гаплар мисолида). филол.фан.бўйича фалсафа доктори (PhD)...дисс. – Тошкент, 2018. –165 б.; Хамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: филол. фан. бўйича фалсафа док. (PhD) ..дисс. – Қарши, 2018. – 250 б.; Эшмўминов А.А. Ўзбек тили миллий корпусининг синоним сўзлар базаси: филол. фан. бўйича фалсафа док. (PhD)...дисс. – Қарши, 2019. – 140 б.; Тонрова Г. Ўзбек тили миллий корпусини яратишнинг назарий ва амалий масалалари. – Германия: Globeedit, 2020. – 168 б.

shakllantirish, tilning leksikografik mashina fondini yaratish kabi qator amaliy ishlar uchun o'rganish obyekti bo'lib xizmat qiladi.

Korpus lingvistikasi turli fanlar uchun o'rganish obyekti va vositasi sifatida foydalanilmoqda. Mazkur sohada Respublikamizning bir qator oliy ta'lim muassasalari, shuningdek, ilmiy tadqiqot institutlarida mazkur yo'nalishlar bo'yicha ilmiy izlanishlar olib borilmoqda. O'zbek korpus lingvistikasida B.Mengliyev, Sh.Shahobiddinova, Z.Xolmanova, S.Karimov, L.Raupova, Sh.Hamroyeva, N.Abduraxmonova, G.Toirova, J.Djumabayeva, G.Ergasheva, A.Eshmo'minov, Sh.Gulyamovalarning tadqiqotlarini qayd etish o'rinli[8].

E'tiboringizga taqdim etilgan Korpus lingvistikasi darsligi orqali kompyuter lingvistikasi doirasida korpus texnologiyasini rivojlantirish bo'yicha olib borilgan ilmiy-nazariy va amaliy natijalar tavsifiga doir ma'lumotlardan bohabar bo'lasiz.

O'qish jarayonida siz quyidagi bilim va ko'nikmalarga ega bo'lasiz:

Korpus lingvistikasida nutqning barcha ifodalarini kuzatish, tahlil qilish, o'rganishning imkoniyati mavjud. Korpus matnlar yig'indisi sifatida o'rganilayotgan obyekt va predmetning tizimli majmuasidir. NLP va kompyuter lingvistikasining asosiy raqamli resursi sifatida korpus til texnologiyasiga doir masalalarni yechishga ko'maklashadi.

Modellashtirishning informatsion, kompyuter, matematik, biologik, raqamli, mantiqiy, statistik, struktur, grafik kabi qator turlari mavjud. O'zbek tili elektron korpusining kompyuter modellari ikki xil yondashuvga asoslanadi: 1) korpus uchun yaratilgan tayyor ingvistik instrumentariylar va platformalar; 2) korpus taksonomiyasidan kelib chiqib, turli maqsadlarga mo'ljallangan kompyuter lingvistikasi metodlari orqali yaratilgan formal-funksional modellar.

---

[8] Mengliyev B., Hamrayeva Sh. Korpus lingvistikasi: korpus tuzish va undan foydalanish. – T.: Globeedit. 2020. – 50 b.; Хамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: филол. фанлари бўйича фалсафа докт. дисс. – Бухоро, 2018. – 165 б.; Mengliev B., Shahabitdinova Sh., Khamroeva Sh., Gulyamova Sh., Botirova A. The morphological analysis and synthesis of word forms in the linguistic analyzer // Linguistica Antverpiensia (1), 2021. – P. 703-712; Тоирова Г. Миллий корпус яратишнинг технологик жараёни хусусида. // Ўзбекистонда хорижий тиллар. Электрон илмий-методик журнал. – Тошкент, 2020. – № 2 (31) – Б.57-64. https://journal.fledu.uz/uz/ 2-31-2020.

Til korpusi davriy silsilaning mahsuli sifatida jamiyat laboratoriyasining o'rganish obyekti vazifasini bajaradi. Shu bois korpusni uzluksiz ravishda boyitib, yangilab borishda korpus konseptologiyasini modellashtirishning universallik, standartlashuv, reprezentativlik kabi mezonlariga asoslaniladi.

Elektron korpuslar lingvistik ma'lumotlarni tahlil qilishda zamonaviy kompyuter metodlari va lingvistik tadqiqotlarni umumlashtirish natijasidir.

Korpus texnologiyasi kompyuter lingvistikasining asosiy obyekti va predmeti sifatida matnlarni qayta ishlashda modeli vazifasini bajaradi. Elektron korpusning tizimlashtirilgan va ma'lum me'yorlarga asoslangan tasnifiy matnlari yordamida lisoniy qoliplar, grammatik qoidalar, lug'atlarni yanada takomillashtirishga hissa qo'shadi. Tilning pragmatik va kognitiv tasviri aks etgan og'zaki va yozma matnlar majmuasi (korpus) ontologik bilimlar bazasi, semantik va neyro to'rlar va sun'iy intellekt texnologiyasi hamda lingvoprotsessor uchun tilning lisoniy modellari va nutqiy aktlarni o'rganishda katta rol o'ynaydi.

Filologiyaning muayyan sohasida ilmiy tadqiqotlar olib borishda maxsus korpuslar yaratishda lingvistik instrumentlaridan foydalanish statistik natijalarga erishishning kvantativ usuli bo'lib xizmat qiladi.

Matnlarni tahlil qilishda muayyan turdagi instrumentariylar korpus foydalanuvchilarning maqsad va vazifalaridan kelib chiqib individual foydalanuvchilar interfeysi va korpus menedjeriga ega bo'ladi.

Parallel matnlar uchun segmentlash va lingvistik tahlil jarayonida Wordfast kabi instrumentlar samarali texnologik vosita sanaladi. Parallel matnlardagi konkordanslarni aniqlashda so'z, so'z birikmasi yoki barqaror birikmalarning u yoki bu tildagi muqobil ekvavalentligini aniqlashda tayyor lingvistik instrumentariylarga tayanish muhim. Parallel matnlardagi kalit so'zlar uchun kontekstda tez-tez qo'llanadigan birliklarni tarjimon xotiraga yuklash orqali parallel matnlarning qidiruv tizimini yaratishda foydalaniladi.

Korpusni lingvistik jihatdan annotatsiyalash (razmetka) undan qay tartibda foydalanishning bosh tamoyili hisoblanadi. Matnlarning razmetkasi uning metama'lumotida (yozma matnlar uchun uning struktur birliklari, sarlavhasi, muallifi, janri, nashri, yili; og'zaki matnlarning obyekti sifatida olingan so'zlovchilarning yoshi, jinsi, kasbi, millati) aks etadi. Kiritilgan ma'lumotlarni mashina o'qiy oladigan formatda matnni kodlash yo'riqnomasi (Text Encoding Initiative (T.E.I.)) orqali standart tilda beriladi.

Lingvistik annotatsiyalash leksik, morfologik, sintaktik, semantik, prosodik (diskurs), anaforik, temporal turlarga tegishli bo'ladi. Annotatsiya jarayonida muayyan xulosani chiqarish uchun annotatsiya modellaridan (annotation scheme) foydalaniladi. Ular annotatsiya qo'llanmalarida qayd etiladi. Matnni annotatsiyalash qo'l mehnati yoki avtomatik va yarim avtomat usulda amalga oshiriladi.

Annotatsiyaning mazmuniy jihatlari yuqorida qayd etilgan lingvistik annotatsiyalash turlaridan tashqari annotatsiya formati belgilsh muhim mezonlardan biridir.

O'zbek tilining morfologik va sintaktik teglash tizimi Protégé texnologiyasi orqali ontologik modellashtirishda iyerarxik munosabatga asoslanilgan. Grammatik teglash tizimi UniTurk doirasida turk, tatar, qozoq va qirg'iz tillari uchun ham assotsiativ shaklda munosabatlar tizimiga bog'langan, natijada korpus analizi uchun turkiy tillarning teglash tizimi yaratilgan.

Chekli avtomat metodi (FST) yordamida o'zbek tilining morfologik analiz qiluvchi dasturning quyida mashina fondi yaratilgan: 1) "Qoidalar-Rules" bunda alifbo, fonologik qoidalar va fonetik hodisaga uchraydigan maxsus fonemalar (bunda o'zbek tili uchun har ikki grafema asos qilib olindi: kirill va lotin); 2) Lug'at-lexicon (barcha so'z turkumlarining o'zak va sodda yasama shakli kiritiladi).

FST texnologiyasiga asoslanish uning ehtimollik variatsiyalarini natija sifatida olishda yordam beradi. FST orqali katta hajmdagi matn korpusi va u yerdan olingan holatlar tizimidan foydalanib, ehtimollik modellar yaratiladi, so'zning imlosini nafaqat tekshirishda, balki uni to'g'irlashda (foydalanuvchiga bir nechta variant taklif qilgan holda) ko'maklashadi.

Matn yetarli miqdorda hajmga ega bo'lishi, turli tipdagi ma'lumotlarni o'zining tabiiy kontekstdagi shaklida korpusda ifodalanishi, tayyorlangan va yaratilgan ma'lumotlardan ko'p marotaba foydalanish imkoniyatining mavjudligi, korpus menejerining foydalanuvchilar uchun olingan natijalarni mos shaklda taqdim qilish, talabdan kelib chiqqan holda korpus funksiyalari imkoniyatalarini yaratish hamda beriladigan so'rovlar natijasida erishiladigan statistik ma'lumotlarni aks ettirish, saralangan va jamlangan matnlarni boshqarish va lingvistik ma'lumotlarni muayyan maqsadlar uchun foydalanishga yo'naltirish korpus yaxlitligini ta'minlaydi.

Lingvistik annotatsiyalangan korpuslar kompyuter lingstikasi yo'nalishlari uchun (mashina tarjimasi, nutq sintezatori, semtiment analiz, spell-cheker kabilar) obyekt vazifasini bajaradi.

Korpusning qidiruv tizimi (menejeri), foydalanuvchilar hamda KL mutaxassislari uchun foydalanish imkoniyatlari bir-biridan farqlanadi. Har ikki subyekt uchun matn xatolarini bartaraf etishda Jaro Vinkler algoritmi va FST texnologiyasidan foydalanish inson resursi sarflaydigan vaqt va mehnatni tejash hamda o'zbek tilining grammatik fondini yaratishda vosita bo'lib xizmat qiladi.

1-mavzu. KORPUS VA KORPUS LINGVISTIKASI

Korpus va korpus lingvistikasiga doir ilmiy qarashlar, tushunchalar va nazariyalar. Fanning maqsadi, predmeti va obyekti. Korpus lingvisikasining yondosh sohalari rivojidagi o'rni va ahamiyati.

XXI asrning global masalalardan biri tabiiy tillarning milliy xususiyatini saqlab qolish hisoblanadi. Dunyo tillarining elektron korpuslarini yaratish va rivojlantirishda tabiiy tilni qayta ishlash (NLP – natural language proccessing) hamda til texnologiyalariga doir tadqiqotlarni izchil rivojlantirish fan oldida turgan muhim vazifalardan biridir. Yo'qolib borish xavfi ostida qolayotgan tillarni saqlab qolish va ularning raqamli resurslarini yaratishga doir ilmiy loyiha va anjumanlar xalqaro tashkilot va jamg'armalar tomonidan qo'llab-quvvatlanmoqda. Shu boisdan tillarning elektron bazasini yaratish millat tili sifatida saqlab qolishning asosiy omili ekanligi soha muatxassislari tomonidan e'tirof etilmoqda. [UNESCO LT4All 2019, TurkLang 2020].

1990-yilgacha olingan ma'lumotlarga ko'ra, dunyo tillarining kompyuter tahliliga mo'ljallangan 600 ga yaqin korpusi borligi aniqlangan[9].

Yillar	Soni
-1965	10
1966-1970	20
1971-1975	30
1976-1980	80
1981-1985	160
1986-1990	320

Ayni vaqtda sun'iy intellekt texnologiyalarining takomillashi natijasida til texnologiyasiga bo'lgan e'tibor va qiziqish yanada ortib bormoqda. Bu borada

[9] Захаров В. П., Богданова С. Ю. Корпусная лингвистика. Учебник для студентов гуманитарных вузов. – Иркутск: ИГЛУ, 2011. – С. 12.

kompyuter texnologiyalari elektron korpus yaratishda tilni qayta ishlash, tahlil qilishning oson, qulay va tezkor vositasi bo'lib xizmat qiladi.

Korpus qachon yaratilgan va uning fan sifatidagi maqomi qanday bo'lgan, degan savol shu sohada izlanish olib borayotgan barcha izlanuvchilarni befarq qoldirmaydi. Korpus lingvistikasining nazariy ildizlari Noam Chomskiy ilgari surgan generativ tilshunoslik bilan bog'liq. Matnlarni to'plam shaklida avtomatik tahlil qilish korpus lingvistikasi taraqqiyoti uchun muhim voqelik bo'ldi.

Manbalarda *kompyuter korpusi* tuchunchasi *korpus lingvistikasi* terminidan avvalroq iste'molga kirgani haqida ma'lumotlar uchraydi. McEnery va Vilson (2001), Tomas Akinas va Alfonsa Yulanda kabi olimlar so'zlarning statistik tahlilini o'rganishda Roberto Busaning ishlariga murojaat qiladi. Ularning fikricha, Braun korpusi "Kompyuterlar va ijtimoiy-gumanitar fanlar" jurnalida ilk korpus sifatida tilga olinadi. Korpus lingvistikasiga doir tadqiqotlarning kengayishi, integrallashgan sohalarda olib borilgan ilmiy hamkorliklar shu yo'nalishda til assosatsiyalari va ilmiy markazlarning tashkil topishida muhim rol o'ynadi.

Graem Kennedi (1998) korpus lingvistikasini soha sifatida shakllanishini Braun korpusi, LOB korpusi va London-Lund korpuslarining yaratilishi bilan bog'laydi. Dastlab tilshunoslar tomonidan ushbu korpuslarga nisbatan tanqidiy fikrlar bildirilgan. 1970-80-yillarga kelib, korpuslarni elektronlashtirish jarayoni ilg'orlashdi. Bunga kompyuter xotirasida katta hajmdagi ma'lumotlarni to'plash, saqlash va uzatish imkoniyatining mavjudligi muayyan darajada rol o'ynadi.

Boshqa bir manbada dastlabki tadqiqot sifatida stenograf Friedrich Wilhelm Kaedingning korpusdan statistik ma'lumotni olish orqali stenografiya tizimini rivojlantirish mumkinligi haqidagi qarashlari asos sifatida keltiriladi. Stenografiyaning optimal tizimini yaratish uchun Kaeding katta hajmdagi loyiha ustida ish olib borgan va bunga 665 ta ko'ngillilar jalb etilgan. Natijada 11 mln.ga yaqin so'zdan 250 mingta so'z va uning sillabuslari ajratib chiqilgan. Loyihaga jalb etilganlar tomonidan so'zlarni aniqlash, hisoblash va saralash ishlari qo'l

mehnati yordamida amalga oshirilgan. Kaeding nemis tilining turli janr va uslubdagi matnlar va ularning reprezentativlik xususiyatlarini o'rganib, ularni kompyuter ma'lumotlar bazasiga kiritgan. Shu orqali nemis tilining turli lisoniy vaziyatlarda qo'llanilishi matn orqali tahlil qilingan. Korpus elektronlashtirilgandan so'ng web sahifalar ham matn (fayl) sifatida kompyuter xotirasiga kiritilgach, ma'lumotlarni boshqarish usullari ishlab chiqilgan.

Katta hajmdagi elektron matnlarni xotirada saqlash imkoniyatining nisbatan cheklanganligi 80-yillarning oxiri 90-yillarning boshida internet (WWW) tarmog'ining yanada kengayib borishi natijasida matematik modellarga asoslangan yangi texnologiyalarni vujudga keltirdi. O'sha davrdan boshlab lingvistikaga ehtimollik modellarning tatbiq etilishi fanga empirizm tushunchasini kirib kelishi va lingvistik grammatikalarning rivojiga ta'sir ko'rsatdi[10].

Aksariyat ilmiy manbalarda dastlabki yozma korpuslar sifatida *Brown* (1964), *LOB (1978), FROWN* (1999), *FLOB (1998), Kolhapur (*1978), *ACE* (1986) manbalar qayd etilsa, nisbatan keyin yaratilgan og'zaki korpuslar sifatida *SEU, LLC, SEC* (ingliz tilining og'zaki matn korpusi), *Map task korpus* hamda *HKCSE (*inglizcha og'zaki matnlarning gong gonk korpusi)[11], *aralash korpus* (og'zaki va yozma matnlardan tashkil topgan) sifatida ingliz tilining xalqaro *ICE* korpuslari e'tirof etiladi. Yuqorida nomlari keltrilgan korpuslarning hajmi 1 mln. so'zni o'z ichiga olgan. Tadqiqotlar uchun bu miqdor yetarli emasligi korpus texnologiyasida tilshunos, dasturchi, matematik, tarjimon hamda pedagoglar hamkorlik olib borish zaruriyatini ko'rsatdi.

Bugungi kunda korpus va elektron korpus terminlari sinonim sifatida ishlatilmoqda[12]. Korpusdan lingvistik tahlil uchun *lingvistik annotatsiya,* olingan

---

[10] Sandra K., Heike Z. Corpus linguistics and linguistically annotated corpora. – New York: Bloomsbury Academic, 2015. – P. 10.

[11] Martin W. Practical Corpus Linguistics An Introduction to Corpus-Based Language Analysis. Oxford, 2016. – P. 16.

[12] Sandra K.Heike Z. Corpus linguistics and linguistically annotated corpora. – New York: Bloomsbury Academic, , 2015. –P. 4.

natijalarning asoslanishi uchun *metama'lumot* terminlari kiritilgan. Tadqiqotimizda korpuslarning funksional imkoniyatlarini alohida tahlil qilamiz.

Tadqiqotchi Martin Weisser matnning hajmiga ko'ra *mega va milliy korpus* terminlarini farqlaydi. Milliy korpus sifatida Britaniya milliy korpusi– BNC (1995), Amerika milliy korpusi – ANC, Zamonaviy Amerika ingliz tili korpusi – COCA, hamda global webga asoslangan korpus – GloWbE, rus tilining milliy korpusi (https://ruscorpora.ru/), Avstraliya milliy korpusi – AusNC turlari mavjud. https://www.english-corpora.org/corpora.asp saytida ingliz tiliga doir bir qancha korpuslar o'rin olgan[13]. Mazkur platformada 22mln. web sahifadan tashkil topgan *iWeb korpus* yaratilgan bo'lib, uning kontenti 14 mlrd. so'zni tashkil qiladi. U *BYU korpuslar* deb nomlangan ingliz tili korpuslari bilan ham bog'liq. Ushbu korpusning asosiy qidiruv tizimi uch usulda amalga oshiriladi: 1) *so'z shakli, so'z turkumi, 60000 so'zning chastotasi ro'yxati va talaffuzi (til o'qituvchilari va o'rganuvchilari uchun); 2) so'z alohida shaklda, kollokatsiyasi, mavzu, tasnifiy, websahifalar, konkordans va har bir so'z uchun ularga bog'liq bo'lgan so'zlar bo'yicha qidiruv tizimi; 3) fraza yoki qator bo'yicha qidirish*. Chunki korpusning ichki qatori qidiruv tizimining tezligi uchun maxsus model, masalan, **ism, un*able* hamda fraza bo'yicha qidirish uchun tilning imkoniyatidan kelib chiqib, formal model ishlab chiqilgan: *got VERB-ed, BUY* ADJ NOUN*. Mazkur korpus ingliz tilining oltita davlatda so'zlashuvchi dialektlar namunalari kiritilgan.

NOW (**N**ews **o**n the **W**eb-Webdagi yangiliklar) korpusi 2010-yildan to shu kungacha bo'lgan web sahifalardagi gazeta va jurnallardan olingan 12,5 mlrd. leksik birlikni o'z ichiga oladi. Mazkur korpusning alohida ahamiyatli jihati shundaki, uning kontenti har oy 180-200 mln. so'z ta'minoti (tahminan 300000 yangi maqolalar) bilan boyitib boriladi. Google Trendsga o'xshagan boshqa resurslar uchun ushbu korpus tilda bo'layotgan o'zgarishlarni tizimga solib turadi. Har yilgi so'zlarning qo'llanilish chastotasi avvalgi yillardan farqli ravishda yangi

[13] https://www.english-corpora.org/

so‘z va frazalarda o‘zgarish kuzatilish mumkin. 1.9 mlrd. so‘zdan iborat *GloWbE* platformasida (**Glo**bal **W**eb-**b**ased English) yigirmaga yaqin davlatda nashr etiladigan yozma manbalardan tashkil topgan.

Hansard korpusi Angliyadagi Lankaster universiteti tomonidan yaratilgan. Mazkur korpus 1803-2005 yillarga oid Britaniya parlamentining 1.6 mlrd. so‘zni qamrab olgan og‘zaki va yozma matnlaridan tashkil topgan bo‘lib, semantik tahlil, lemmatizatsiya, morfologik va semantik teglash kabi funksiyalarga ega.

Zamonaviy Amerika ingliz til korpusi (*COCA*) matn reprezentativligiga ko‘ra teng taqsimlangan, katta hajmli korpus sanaladi. Ushbu korpus tarkibiga 1990-2019-yillarni qamrab olgan sakkiz xil janrga tegishli matnlar kiritilgan (og‘zaki matn, badiiy matn, ommabop jurnallar, gazetalar, ilmiy matnlar). Mazkur korpusga 25 milliondan ortiq web sahifalardan olingan ma’lumotlar yuklangan. COCA korpusi ham *Iweb korpusi* singari beshta usulga asoslangan qidiruv imkoniyatlariga ega.

Rus tili korpusining arxitekturasi ikki massivli yozma matnlardan iborat. Korpusning birinchi qismi XX asr o‘rtalari va XXI asr boshlarida yaratilgan matnlar, ikkinchi qismi XVIII asr oxiri- XX asr boshlariga tegishli yozma manbalarni o‘z ichiga olgan. Uning kontentida rus tilining yozma adabiy tiliga doir zamonaviy badiiy proza janrlari va uning yo‘nalishlari, zamonaviy dramaturgiya, memuar va biografik adabiyotlar, jurnalistika va adabiy tanqid, gazeta jurnalistikasi va yangiliklar, ilmiy, ilmiy-ommabop va o‘quv materiallar, diniy va diniy-falsafiy matnlar, rasmiy ish va qonun hujjatlari, kundalik matnlar (shu jumladan nashr uchun mo‘ljallanmagan matnlar: shaxsiy yozishmalar, kundaliklar va boshqalar) o‘rin olgan.

Rus tili milliy korpusida matnlarni turli jihatdan annotatsiyalash mumkin bo‘lgan bir nechta bo‘limlar mavjud[14]. SinTagRus (Syntactically Tagged Russian corpus) nomli chuqur annotatsiyalangan korpus morfosintaktik razmetkadan iborat.

[14] https://ruscorpora.ru/new/corpora-structure.html

Razmetkalashda I.A.Melchuk va A.K.Jolkovskiyning "Mazmun⇔Matn" konsepsiyasiga asoslangan shajarasimon tahlil metodiga tayanilgan. Rus tilining milliy korpusi rus tili gazetalar matni, parallel korpus, hududiy va xorijiy nashrlar, poetik korpus, dialektlar korpusi, ta'limiy korpus, og'zaki matn korpusi, aksentologik korpus, multimediyali kabi korpuslardan iborat.

Ma'lumki, nutqning yozma va og'zaki shakllari mavjud bo'lib, uning xoslik belgisi kishilar muloqoti orqali turli uslublardagi matn mazmunida o'z aksini topadi. Tilning bu imkoniyati nutqiy vaziyatda uning cheksizligini ko'rsatadi. Dunyoda nechta til bo'lsa, u shuncha millat va elat nutqida pragmatik, lingvomadaniy ko'rinishlarga ega. So'zlovchilar nutqining turli omillar, chunonchi, obyekt va subyekt munosabatiga ko'ra farqlanishi korpus orqali namoyon bo'ladi. Korpus lingvistikasida nutqning barcha ifodalarini kuzatish, tahlil qilish, o'rganishning imkoniyati mavjud. Korpus matnlar majmui sifatida o'rganilayotgan obyekt va predmetning tizimli majmuasidir. Bu borada V.Zaxarov: *korpus tilning qisqartirilgan modeli,* deb o'rinli baho beradi. Zero, tilning turli diskursdagi ifodasi tabiiy nutq holatida voqelanadi.

Ch.Taylor "Korpus lingvistikasi nima?" sarlavhali maqolasida[15] korpus lingvistikasi hamisha chastotaga asoslanishi, undagi ma'lumotlar veb sahifalaridagi matnlardan tashqari nutqiy jarayonda yaratilgan turli uslubdagi yozma va og'zaki materiallar (gazeta va jurnal materiallari) hamda audio ko'rinishidagi ma'lumotlardan tashkil topishi bilan farqlanishini o'z kuzatishlari orqali ilmiy asoslashga harakat qiladi. Shu jihatdan tadqiqotchi korpusning umumiy va maxsus turlarga ajratilishi, maxsus turdagi korpuslar janr, uslub, davrlarga ko'ra farqlanishi, har ikki turdagi korpus o'z navbatida diaxron va sinxron shaklida bo'lishi mumkinligini qayd etadi. Shuningdek, olim o'zining ilmiy qarashlarida diaxron va sinxron korpuslar haqidagi fikrlarini bildiradi. Unga ko'ra diaxron korpus tilning davrlar osha qay darajada o'zgarganligini glottoxronologik va

[15] Charlotte T. What is corpus linguistics? What the data says // ICAME Journal, 2008. – P. 32.

statistik tahlil qilish uchun obyekt vazifasini bajarsa, sinxron korpus nutqiy birliklarning ayni vaqtda qo'llanilayotgan zamonaviy ko'rinishlarini ifodalaydi. Shuningdek, Ch.Taylor izlanishlarida korpus turli soha vakillari uchun maxsus o'rganish vositasi sifatida *monolingval* va *parallel* matnlarning majmuasidan iborat mumkinligini asarlarida qayd etadi.

Ch.Mirning fikricha, korpus izlanish olib borilayotgan tor sohaga doir tadqiqotning metodologik usulini ham korpus deb nomlash mumkin. Olim tomonidan quyidagi munozarali savol beriladi: maqolalarning onlayn korpusi yaratilsa, elektron shaklidagi barcha maqolalar uning korpusimi yoki matnlar korpusi tahlilidan hosil bo'lgan maqolalar uning tahlil natijasimi? U korpusga ta'rif berishda til injeneriyasi sohasining standartlari bo'yicha ekspertlar maslahat guruhi (The Expert Advisory Group on Language Engineering Standards – EAGLES) tomonidan korpusga berilgan ta'rifni asarida qayd etadi. Unga ko'ra, nafaqat nasr, gazeta nashrlari, nazm, dramaga oid matn turlari, balki so'zlar ro'yxati yoki lug'atlar ham korpus bo'lishi mumkin[16]. Ushbu ta'rifga ko'ra, barcha lingvistik manbalar korpus sifatida baholangan[17]. Korpus matnlarning jamlanmasidir, degan fikrga olim o'z munosabatini bildiradi hamda Otto Jespersning ko'p seriyali tarixiy tamoyillar asosidagi "Zamonaviy ingliz tili grammatikasi"da matnlarning turli lingvistik strukturalari Chauser, Shekspir, Svift, Austin va Jespersenlarning asarlaridan olingan matn namunalarini keltiradi. Biroq bu korpusga asoslangan tahlil deyilmasa-da, uni matn fragmentlaridan pedagogik va leksikografik tadqiqotlar uchun foydalanilishi matnlar majmuasining korpus analogi sifatida baho berilishiga turtki bo'lgan.

N.Chomskiyning tomonidan korpusga berilgan ta'riflarda shunday deyiladi: til chekli qoidalar va me'yorlarga ega va undan cheksiz jumlalar yaratish imkoniyati mavjud. Biroq til jamiyatda turli ijtimoiy omillar ta'sirida, neologizmlarning kirib kelishi korpusning nechog'lik katta hajmga ega bo'lmasin

[16] "Corpus Encoding Standard": http://www.cs.vassar.edu/CES/CES1-0.html10
[17] Charlez M. English corpus linguistics: An introduction. Cambridge University Press, 2004. –P. 168.

tilda aks etuvchi barcha nutqiy imkoniyatlarni bir vaqtning o'zida jamlay olmasligini qayd etadi. Bizningcha, olimning matn korpusiga bildirgan fikrlari ayni haqiqat. Zero, korpus har qancha ulkan bo'lmasin tilning yaxlit manzarasini tasvirlashga qodir emas. Demak, istalgan matn korpusi barcha tadqiqotlar uchun universal bo'la olmaydi.

Ta'kidlanganidek, korpus yoki korpus tilshunosligidagi olimlarning nuqtai nazarlari turlicha. Korpusni til modeli emas, metodologik yondashuv sifatida baholash Dougles Biberning quyidagi fikrlarida tahlil qilingan[18]. Korpus:

• tabiiy matndagi kerakli birliklarni empirik tahlil qiladi;

• tahlil uchun "korpus" sifatida tabiiy matnlarning katta va tizimlashtirilgan jamlanmalarini birlashtiradi;

• tahlil uchun kompyuterning ham avtomatik va interaktiv texnologiyalaridan foydalanish imkonini beradi;

• analitik texnologiyaning miqdor (statistik) va sifat xususiyatlarini o'z ichiga oladi.

Til davriy silsilaning mahsuli sifatida jamiyat laboratoriyasining o'rganish obyekti hisoblanadi. Shu bois og'zaki va yozma nutqdagi uslubiy xoslanish, qolaversa, til va jamiyat munosabati ta'sirida tilning dinamik xususiyati korpusga bo'lgan ehtiyojni yuzaga keltirishi olimlar tomonidan e'tirof etiladi. Natijada korpusni uzluksiz ravishda boyitib, yangilab borish ehtiyoji mavjud. Bu jihat korpusshunoslik sohasida lingvistik tadqiqotlarning izchil va uzluksiz rivojlantirish zaruriyatini belgilaydi.

Tillarning lingvistik resurslar bazasini yaratishda standart modellarga asoslanilgan elektron korpusning ahamiyati katta. Turkologiyada bu borada muayyan darajada tajribalar to'plangan: turk tilining milliy korpusi (TNC) - www.tnc.org.tr ; qozoq tilining Almaty Corpus (NCKL) - veb - corpora.net/KazakhCorpus; oltoy tilining korpusi - altay2.gasu.ru ; boshqird

[18] Bern H., Heiko N. The Oxford Handbook of Linguistic Analysis. / Dougles Biber Corpus-based and Corpus-driven analysis of language variation and use UK: Oxford university, 2015. –P. 193.

tilining milliy korpusi - bashcorpus.ru ; boshqird tilining poetik korpusi - web-corpora.net/bashcorpus; tatar tilining milliy «Tugan Tel» - tugantel.tatar ; tatar tilining yozma matnlar korpusi - www.corpus.tatar; xakas tilining korpusi - khakas.altaica.ru; yoqut tilining korpusi - adictakha.nsu.ru/corpora/corp; sibir ozchilik tillarining raqamli korpusi (Teleut va Shor) - corpora.iea.ras.ru/corpora.

Turkologiyada korpusshunoslik (korpus lingvistikasi) sohasi bo'yicha quyidagi olimlarning ishlari diqqatga sazovor: turk tili bo'yicha M.Aksan, D.Zeyrek, K.Oflazer, U.Özge Bular, Eshref Adali; uyg'ur tili bo'yicha Y.Aibaidulla, K.Lua; boshqird tili bo'yicha L.A.Buskunbaeva, Z.Sirazitdinov[19]; hakas tili bo'yicha A.Sheimovich[20], tatar tili bo'yicha J.Suleymanov, A.Gatiatullin, O.Nevzorova, R.Gilmullin, B.Hakimov; qirimtatar tili bo'yicha L.Kubedinova hamda tuva tili bo'yicha A.Salchak kabilar.

Tatar tilining "Tugen tel" deb nomlangan milliy korpusi Amaliy semiotika instituti tomonidan (2012-2014) yaratilgan. Qayd etilishicha[21], yuqorida sanab o'tilgan tillar orasida tatar tilining korpusi barcha til sathlarining annotatsiyalangani va qo'yilgan talablarga nisbatan to'liq javob berishi bilan boshqa turkiy korpuslarga qaraganda ajralib turadi. Tatar tilining milliy korpusi barcha lingvistik mezonlarga moslashtirilgan konseptual va funksional modellar majmuasidan iborat. Konseptual va funksional modellarning sinflari o'z navbatida, muayyan til sathining struktural va funksional tavsifi, shuningdek, tabiiy tilni qayta ishlash uchun til texnologiyasi va axborot tizimlarini rivojlantirish uchun zarur bo'lgan umumiy ma'lumotlardan tshkil topadi.

Turkiy tillar bo'yicha korpus sohasiga doir qator tadqiqotlar amalga oshirilgan. Manbada qayd etilishicha[22], turk tilining 423 mln. so'z, 491 tokendan

---

[19] Бускунбаева Л. А., Сиразитдинов З. А. К системе разметок в национальном корпусе башкирского языка // Актуальные проблемы диалектологии языков народов России. Материалы XI межрегиональной конференции. –Уфа, 2011. – С. 50–55.

[20] Шеймович А. В. Морфологическая разметка корпуса хакасского языка // Российская тюркология, 2011.– № 2(5). – С. 48–61.

[21] Suleymanov D. et al. National Corpus of the Tatar L Grammatical Annotation and Implementation // 5th International Conference on Corpus Linguistics (CILC2013), 2011. – P. 68-74.

[22] https://www.sketchengine.eu/wp-content/uploads/Large_Corpora_for_turkic_2012.pdf

iborat BOUN korpusi, 2 mln. so'zdan iborat METU, 50 mln. so'zli web korpusi 34 Mbdan iborat parallel korpus kontentidan iborat. Olimlar turk, ozarbayjon, o'zbek, qozoq, turkman tillarining korpusini morfologik tahlil etishda matnlarni segmentlash orqali turkiy tillarning morfotaktik modellarini ishlab chiqdilar. Morfologik tahlil algoritmi morfemalarning lisoniy model chegarasiga ko'ra asoslanib, kichik hajmdagi lug'atdan to 70 ming hajmdagi turkcha so'zlarning morfologik shakllarini qamrab olgan.

Elektron lingvistik korpuslarni rivojlantirishda lingvistik ma'lumotlarni tahlil qilish uchun zamonaviy kompyuter metodlari va lingvistik tadqiqotlar natijalarini birlashtirish muayyan darajada muammolarni yechishning optimal usuli hisoblanadi. Korpuslardagi annotatsiyalash tizimining standartlashuvi ulkan hajmdagi ma'lumotlarni qayta ishlash imkonini yaratadi.

Ilmiy faoliyatning sohalararo integratsiyasi natijasida bir-biriga yaqin tillarning korpuslarida aks etuvchi grammatik kategoriyalarni teglash tizimi hamda umumiy annotatsiyalash jarayonidagi muayyan darajada o'xshashlik korpuslarning lingvistik reprezentativligini ta'minlashga hizmat qiladi. Barcha turkiy tillar doirasida matnlarni lingvistik annotatsiyalash tizimi uchun umumiy grammatik teglash va annotatsiyalash tamoyillari hamda mezonlari ishlab chiqilsa, tabiiy matnlarni qayta ishlashning ko'p tilli texnologiyalarida foydali model bo'lib xizmat qilishi, shubhasiz.

Mavzu yuzasidan savol va topshiriqlar:

1. Korpus lingvistikasining obyekti va predmetiga nimalar kiradi?
2. Korpus lingvistikasining soha sifatida shakllanishi va olib borilgan tadqiqotlar bo'yicha fikringizni bayon eting.
3. Korpusga oid saytlar bilan tanishib, quyidagi jadvalni to'ldirishga harakat qiling hamda qiyosiy tahlil qiling.

№	Sayt manzili	Til	Funksional imkoniyat	Yaratilgan sanasi	Subkorpuslari yoki kontenti
1.					
...					

4. Korpus lingvistikasining bugungi tilshunoslikdagi muammolarni o'rganishda qanday amaliy ahamiyati bor? Misollar keltiring.
5. Korpus bo'yicha turli yondashuvlar mavjudligi bois quyidagi jadvalda fikringizni dalillash uchun kamida ikkitada misol yozing.

A.	**Korpus =>metod**	**Chunki,**	**Masalan,**
B.	Korpus=> metodologik fan	Chunki,	Masalan,
C.	Korpus=>instrument	Chunki,	Masalan,
D.	Korpus=> lingvistik resurs	Chunki,	Masalan,

KAZUS

1. Ushbu fikr kimga tegishli? "Til chekli qoidalar va me'yorlarga ega va undan cheksiz jumlalar yaratish imkoniyati mavjud. Biroq til jamiyatda turli ijtimoiy omillar ta'sirida, neologizmlarning kirib kelishi korpusning nechog'lik katta hajmga ega bo'lmasin tilda aks etuvchi barcha nutqiy imkoniyatlarni bir vaqtning o'zida jamlay olmaydi?"	a. *D.Biber* b. *N.Chomskiy* c. *Ch.Meyer* d. *V.Zaxarov* e. *Mcenery* f. *H.Bern*
2. Korpusga berilgan ushbu fikr	

kimga tegishli? "Korpus nafaqat nasr, gazeta nashrlari, nazm, dramaga oid matn turlari, balki so'zlar ro'yxati yoki lug'atlar ham korpus bo'lishi mumkin"	
3. Korpus izlanish olib borilayotgan tor sohaga doir tadqiqotning metodologik usul ekanligi qaysi olim tadqiqotlarida aks etgan?	
Korpus tilning qisqartirilgan modeli, deya ta'rif bergan olimni aniqlang?	

Mavzu yuzasidan test.

1. Kaeding tomonidan amalga oshirilgan korpus yaratish loyihasi qaysi til uchun mo'ljallangan edi ?
   a) *Nemis tili
   b) Ingiliz tili
   c) Fransuz tili
   d) Ispan tili
2. Dastlabki yozma korpuslar berilgan qatorni toping.
   a) Brown, FLOB, LLC.
   b) Brown, LOB, ICE.
   c) *Brown, LOB, ACE.
   d) Brown, FLOB, ICE.
3. " Korpus lingvistikasi hamisha chastotaga asoslanadi" degan g'oyani ilgari surgan olim qaysi javobda ko'rsatilgan ?
   a) V.Zaxarov
   b) * Ch.Taylor
   c) N.Chomskiy
   d) I.A.Melchuk
4. **V.Zaxarovning** korpus haqida bildirgan tarifi qaysi qatorda berilgan.
   a) "Bilimlar yoki dalillar to'plami".
   b) * "Korpus tilning qisqartirilgan modeli".
   c) "Korpus ma'lum bir tildagi ilmiy va badiiy matnlar to'plami".
   d) "Korpus tabiiy tilning qayta ishlangan ko'rinishi".
5. **COCA** korpusida necha xil usulda qidirish imkoniyati mavjud ?
   a) 3 xil
   b) 6 xil
   c) *5 xil
   d) 4 xil
6. " Monitor korpus" tushunchasi fanga kim tomonidan kiritilgan ?
   a) N.Chomskiy

b) Ch.Taylor
c) I.A.Melchuk
d) *J.Senkleyer

7. ... olingan tematik mavzu teng miqdorda taqsimlanadi.
a) *Muvozanatlashgan korpusda.*
b) *Piramidali korpusda .*
c) *Imkoniyat darajasidagi korpusda.*
d) *Biror bir sohaga yo'naltirilgan korpusda.*

8. Zamonaviy Amerika ingiliz tili korpusi qaysi qatorda berilgan?
a) *COCA
b) CONCA
c) COHA
d) OEC

9. Bolalar tiliga xos og'zaki matnlar transkripsiyasi keltirilgan korpus bu...
a) *CHILDES
b) CHILDNESS
c) CHILDS
d) FORCHILD

10. Tillarning elektron bazasini "Korpus" ini yaratishdan asosiy maqsad nima? Til o'qitishda manba sifatida foydalanish uchun.
a) Ilmiy tadqiqotlarda elektron baza sifatida foydalanish uchun.
b) Raqamli texnologiyalar asrida ma'lum bir tilning boshqa tillardan ortda qolmasligini, yo'qolib ketish xavfi ostidagi qo'lyozmalarning umrboqiyligini ta'minlash uchun.
c) *Ma'lum bir tilning millat tili sifatida saqlanib qolishini ta'minlash uchun

11. Kam ta'minlangan resursli tillar guruhini belgilang.
a) Xom-som tillar guruhi
b) *Turkiy tillar guruhi
c) Hind-yevropa tillar guruhi
d) Roman-german tillar guruhi

12. 1990 - yilda dunyo tillarining kompyuter tahliliga mo'ljallangan korpuslar soni to'g'ri keltirilgan qatorni belgilang .
a) *600 ga yaqin
b) 320 ta
c) 540 dan ziyod
d) 160 ta

1.	**Tugen Tel**		Tatar tilining milliy korpusi
2.	arxiv	archive	Ko'p hollarda, korpusga o'xshatilsa-da, o'ziga

			xos tafovutlari mavjud. Jeofri Lichning (1991:11) ta'kidlashicha:"Arxiv va korpus o'rtasidagi asosiy farq shundaki, korpus ma'lum bir "funksiyaga" yo'naltirilgan holda ishlab chiqiladi. Arxiv esa shunchaki "katta hajmli ma'lumotlarning imkoniyat darajasida bir joyga to'plangan va to'la tartibga solinmagan to'plami"dir (Kennedi, 1998:4).
3.	Britaniya Ilmiy Og'zaki Ingliz tili Korpusi (BIOI)	British Academic Spoken English Corpus	Ushbu korpus Amerikaning MIIO (Michigan Ingliz tilining Ilmiy Og'zaki Korpusi) korpusiga qo'shimcha tarzida Britaniyaning Varvik va Riding universitetlari tomonidan ishlab chiqilgan. Bu audio va yozma tarzdagi, katta hajmli, ma'ruza va seminar videotasmalaridan iborat matnni kodlashtirish uskunasidir. Ushbu korpus to'rtta akademik domenlarni o'z ichiga olgan freym asosida shakllantirilgan: San'at va ijtimoiy fanlar, Jamiyatshunoslik, Jismoniy tarbiya va Tibbiyot fanlari. Korpusga oid qo'shimcha ma'lumotlar uchun: http://www.rgd.ac.uk/AcaDepts/ll/base_corpus elektron manziliga murojaat qilishingiz mumkin.
4.	Britaniya Milliy Korpusi (BMK)	British National Corpus	90 foiz yozma, 10 foiz og'zaki shakldagi jami 100 millionta ingliz cha so'zni o'z ichiga olgan korpus. 4, 124 ta matn 1980-1990-yillarga tegishli bo'lsa-da, 5.5 million so'z birinchi marta 1960-1984-yillarga kelibgina chop etilgan. Yozma matnlar tuman va viloyat gazetalari, yilnomalar, barcha yosh va qiziqishdagi insonlar uchun mo'ljallangan jurnallar, akademik kitoblar, mashhur badiiy asarlar, nashr etilgan va etilmagan maktublar, memorandumlar, maktab va universitet inhsolaridan iborat. Og'zaki qismi esa turli yoshdagi, shevaga va qatlamga ega, ko'ngillarning nutqlari va rasmiy matnalardan tortib hukumat uchrashuvlari-yu radio ko'rsatuvlargacha bo'lgan og'zaki matnlardan iborat. Korpusda ma'lumotlar CLAWS CS tegsetidan foydalanilgan holda so'z turkumi

			bo'yicha teglangan. Loyiha Oksford Universiteti Nashriyoti tomonidan amalga oshirilgan va Oksford hamda Lankaster Universitetlari, Eddison-Uesli Longman va Larus Kingfisher Chambers nashriyot uylarining tajribalaridan ham foydalanilgan.
5.	korpus	Corpus (pl. corpora)	lotin tilidan olingan bo'lib ko'pligi **korpuslar (corpora)**. Tilshunoslikda korpus deb, electron ma'lumotlar bazasiga yig'ilgan matnlar to'plamiga aytiladi. Korpuslar esa asosan ming yoki millionlab so'zlardan iborat kompyuter o'qiy oladigan matnlardan iborat katta to'plamlar yig'indisi. Korpus ba'zan arxivdan o'zining muayyan bir tilning farqlliliklarini yoki janrlarini yoki standartlarini ko'rsatish uchun tanlab olingan matnlardan iborat bo'lganligi bilan farqlanadi (lekin doim ham emas). Korpuslar odatda so'z turkumlari teglari yoki nutq bilan bog'liq prozaviy xususiyatlarni aniqlovchi qo'shimcha annotatsiyalarni o'z ichiga oladi. Alohida matnlarning janri, muallifi, nashr etilgan sanasi va joyi haqida ma'lumot beruvchi meta-kodlari bo'ladi. Korpuslarning ixtisoslashgan, havolali (reference), ko'p tilli, parallel, o'rganuvchi, diaxron va monitor kabi turlari mavjud. Ulardan ham sifat ham sonni tahlil qilishda foydalanish mumkin. korpus til haqida yangi ma'lumotga ega bo'lmasada, ma'lumotlar bilan ishlovchi dasturiy ta'minotdan foydalanib, yangi qarashlarga guvoh bo'lishimiz mumkin (Ganston 2002: 2-3).
6.	**Korpusga asoslangan**	**(corpus-based)**	Tognini-Bonelli 2001-yilda korpusga asoslangan va korpus yordamidagi analizlar o'rtasidagi muhim farqni ko'rsatib berdi. Birinchi tur korpusdan tadqiqotchi xohlagan narsani tekshirish uchun misollar manbai sifatida yoki kichikroq ma'lumot setidagi til ishonchliligi va chastotasini tekshirish uchun

			ishlatiladi. Tadqiqotchi oldindan mavjud bo'lgan an'anaviy tavsiflovchi birliklar va turkumlarni tekshirmaydi. Korpus yordamidagi analiz induktiv jarayon bo'lib, korpusning o'zi bir ma'lumot hisoblanadi va undagi qolip/shakllar tildagi odatiylik (yoki ishonchlilik)ni ifodalash usuli sifatida beriladi. Korpus yordamidagi analizda faqat grammatik structuralar haqidagii kichik nazariy faqarlardan foydalaniladi.
7.	**Diaxron korpus**	**diachronic corpus**	til va tildagi o'zgarishlarni ma'lum vaqt davomidagi holatni namoyon etadigan tartibda tuzilgan korpus bo'lib, u yordamida tadqiqotchilar tilga oid o'zgarishlarni yozib borishlari mumkin. Masalan, **Helsinki Corpus of English Texts: Diachronic Part** m.a. 750-yildan to milodiy 1700-yilga qadar davrni qamrab oluvchi 400 matndan tuzilgan.

Adabiyotlar:

1. Abduraxmonova N. O'zbek tili elektron korpusining kompyuter modellari. Globedit, 2021. 220 b.
2. Bern H., Heiko N. The Oxford Handbook of Linguistic Analysis. / Dougles Biber Corpus-based and Corpus-driven analysis of language variation and use UK: Oxford university, 2015. –P. 193.
3. Charlez M. English corpus linguistics: An introduction. Cambridge University Press, 2004. –P. 168.
4. Charlotte T. What is corpus linguistics? What the data says // ICAME Journal, 2008.
5. https://www.sketchengine.eu/wp-content/uploads/Large_Corpora_for_turkic_2012.pdf
6. Martin W. Practical Corpus Linguistics An Introduction to Corpus-Based Language Analysis. Oxford, 2016.
7. Sandra K., Heike Z. Corpus linguistics and linguistically annotated corpora. – New York: Bloomsbury Academic, 2015.
8. Suleymanov D. et al. National Corpus of the Tatar L Grammatical Annotation and Implementation // 5th International Conference on Corpus Linguistics (CILC2013), 2011. – P. 68-74.
9. Бускунбаева Л. А., Сиразитдинов З. А. К системе разметок в национальном корпусе башкирского языка // Актуальные проблемы

диалектологии языков народов России. Материалы XI межрегиональной конференции. –Уфа, 2011. – С. 50–55.
10. Захаров В. П., Богданова С. Ю. Корпусная лингвистика. Учебник для студентов гуманитарных вузов. – Иркутск: ИГЛУ, 2011. – С. 12.

2-mavzu. KORPUS LINGVISTIKASINING SHAKLLANISH TARAQQIYOTI

Korpus texnologiyasining fan sifatida shakllanishi va rivojlanish bosqichlari. Korpus lingvistikasiga doir ilmiy qarashlar

Til hamisha kishilik jamiyatning eng ajralmas aloqa vositasi bo'lib xizmat qilgan. Uning yordamida yangi bilimga erishiladi va egallangan bilim tafakkur va tajriba asosida qayta shakllanadi. Bugun til sanoatida kompyuter lingvistikasi, mashina tarjimasi, til texnologiyasi, tabiiy tilni qayta ishlash, sun'iy intellekt texnologiyasi kabi terminlar tez-tez qo'llanilib kelinmoqda. Bularning har biri insoniyat tomonidan yaratilgan lingvistik resurs orqali o'z tadrijiy takomiliga ega. Endilikda lingvistik resurslarning asosiy turlaridan biri sifatida elektron shaklidagi korpuslar nazarda tutilmoqda.

Dastlabki lingvistik korpus o'tgan asrning 60-yillarda paydo bo'lgan. 1963-yilda Braun universitetida (AQSh) birinchi marta mashina muhitida matnlarning katta korpusi yaratilgan (Brown corpus). Korpus mualliflari V. Frensis va X. Kucher hisoblanadi. Ushbu korpusga turdagi janrlardan tashkil topgan nasriy matnlardan iborat bo'lgan. Barun korpusining ommallashuvi shu sohaga doir olimlar orasida turli muhokamalarni yuzaga keltirdi.

90-yillarning birinchi yarmida korpus lingvistikasi til haqidagi alohida soha sifatida shakllanib ulgurdi. Shu bilan birga, u kompyuter lingvistikasi bilan birga bilan chambarchas aloqada bo'lib, uning yutuqlaridan foydalanadi va o'z navbatida kompyuter lingvistikasi uchun lingvistik resurs sifatida turli dasturiy ta'minotlar uchun tadqiqot obyekti bo'lib hiosblanadi.

Korpus lingvistikasi matnlarni yig'ish, tasniflash, annotatsiyalash kabi vazifalarni bajaradi. Korpus mashina uchun tabiiy tilni tushunish jarayonini lingvistik jihatdan aniq va to'g'ri ko'rsata olishi bois tabiiy tilni qayta ishlash (NLP-natural language processing) sohasida uning o'rni muhim ahamiyatga ega. Shu bois ilmiy manbalarda korpus lingvistikasi tabiiy tilni qayta ishlash sohasining

(NLP) bir yo'nalishi sifatida qayd etiladi[23]. Ayrim ma'lumotlarda[24] esa kompyuter lingvistikasi yoki amaliy lingvistikaning muayyan sohasi deb qaraladi.

Korpus lingvistikasini kompyuter lingvistikasining asosiy yo'nalishi sifatida e'tirof etish mumkin, chunki matnni qayta ishlovchi dasturlar korpusga asoslanadi. NLP tizimi uchun lingvistik bilimlar juda muhim. Ular qabul qilingan muayyan lisoniy modellar, qoidalar, lug'atlar shaklida ifoda etilsada, biroq tilning nutqiy omili bilan bog'liq bo'lgan ekstralingvistik ma'lumotlarga ham ehtiyoji katta. Shu ma'noda tilning konseptual tasviri aks etgan semantik tarmoqlar, ontologiya ko'rinishidagi bilimlar ta'minoti zarur. Demak, tilning murakkab funksional imkoniyatlarini pragmatik, neyrolingvistik, psixolingvistik yoki diskursiv parametrlarini baholashda korpus texnologiyasi oldida muhim vazifalar mavjud.

Kaliforniya universiteti mutaxassisi Stefan Gries[25] o'zining ilmiy qarashlarida korpus uchun berilgan ta'riflar "*u metod, nazariya yoki modelmi*", degan savolga nisbatan uni metod(ologiya) deb baholaydi. Olim fikrini asoslab, til nazariyasining generativ nuqtayi nazaridan tilga aloqasi yo'q, degan to'xtamga keladi. U deskriptiv va amaliy jihatdan ayrim metodlardan foydalanib, misollar yordamida o'z fikrini isbotlashga harakat qiladi. Unga ko'ra, agar tilshunos biror leksemani korpus ichida grammatik birlik sifatida o'rganmoqchi bo'lsa, demak u grammatik nazariyaga, agar u yoki bu millatning ikkinchi til sifatida murakkab konstruksiyalardan qay darajada qo'llash imkoniyatini baholamoqchi bo'lsa, ikkinchi tilni o'rganish nazariyasiga asoslanadi, ya'ni har bir izlanuvchi o'ziga xos metodni tanlaydi, degan munosabatni bildiradi.

Bugungi kunda korpusning imkoniyatlari shu darajaga yetdiki, ushbu sohada erishilgan yutuqlardan nafaqat NLP sohasi yoki kompyuter lingvistikasi, balki tilga o'qitishning metodika va pedagogika sohalari, diskurs tahlil, mashina tarjimasi va

---

[23] Mohamed Zakaria Kurdi Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax. – Great Britain: Wiley-ISTE, 2016. – P. 12.

[24] Jurafskiy D., James H. Martin Speech and language processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, – USA, 2007. – P. 26.

[25] Gries S. Th. How to use statistics in quantitative corpus analysis. In Michael McCarthy & Anne O'Keeffe (eds.), The Routledge Handbook of Corpus Linguistics. – New York, London: Routledge., 2021.

lingvistikaning o'nlab sohalari (sotsiolingvistika, gendershunoslik, psixolingvistika va h.k.) unumli foydalanib, sohaga doir ijobiy natijalarga erishib kelinmoqda. Shu bilan birga Jurafsky, Martin, Manning, Schütze, Roark, Sproat kabi olimlar nuqtayi nazaricha kompyuter lingvistikasi matnlar korpusini annotatsiyalash va uni lingvistik tahlil qilishda o'zining optimal yechimlarini turli algoritmlar misolida tavsiya qiladi[26].

Korpus yozma va transkripsiya qilingan og'zaki nutqni lingvistik tahlil etishda asos vazifasini bajaradi. G.Kennedining fikricha[27], korpus lingvistikasi kompyuterlar rivojlanish davri bilan boshlanmagan bo'lsa-da, matnga asoslangan lingvistik tahlil uchun ma'lumotlar bazasi kompyuterlar keskin ravishda o'sishiga sabab bo'ldi. Korpus turli maqsadlardan kelib chiqib, uning dizayni, hajmi hamda har bir korpusning individual xarakterini shakllantirishga xizmat qiladi. Ko'pchilik lingvistik korpuslar tilning turli leksik, grammatik, diskurs va pragmatik jihatdan tahlil qilish uchun yo'naltiriladi. Jumladan, maxsus maqsadlarga yo'naltirilgan korpuslarda o'quv lug'atida qaysi so'zlar bo'lishi va ularning ma'nolari qaysi konteksda xususiylashuvi aks etadi. Shuningdek, korpus jamiyatning muayyan soha vakillari tomonidan qo'llaniladigan lug'at chastotasini aniqlash, tilning aktiv va passiv lug'atini nazorat qilish kabi vazifalarni bajarishga yordam beradi.

Kompyuter lingvistikasida korpus muhim ahamiyatga ega, ayniqsa, elektron lug'atlar bazasi hamda grammatik qoidalarni yaratishda matnlar majmuasi (korpus)dan foydalaniladi. Korpus kompyuter lingvistikasining mashina tarjimasi, nutq sintezatori, matn tahlili, sentiment tahlil va boshqa qator yo'nalishlar uchun tadqiqot obyekti vazifasini o'taydi.

Amerikalik filolog Georg Zif matnda qo'llangan so'zlar chastotasi bilan matn uzunligi o'rtasidagi munosabatlarga asoslangan Zif qonunini yaratdi[28].

---

[26] Jurafsky, D., Martin J. Speech and language processing, 2nd ed. Upper Saddle River: Prentice Hall, 2008; Manning C., Schütze H. Foundations of statistical natural language processing. – Cambridge: The MIT Press, 1999; Roark, B., Sproat R. Computational approaches to morphology and syntax. –Oxford: Oxford University Press, 2007; Mitkov R. The Oxford handbook of computational linguistics. – Oxford, 2003. – P. 63

[27] Graeme K. An introduction to corpus linguistics. – London: Longman, 1998. – P. 2

[28] O'sha joyda. – P. 10

Shuningdek, lingvistik tadqiqotlarda matnga asoslangan yondashuv Jorjtaun universitetida bo'lib o'tgan muhokamada o'z aksini topdi. Unga ko'ra, korpusning quyidagi yo'nalishlari bo'yicha rivojlantirishga e'tibor qaratildi[29]: 1) og'zaki matn korpusini loyihalashtirish va takomillashtirish; 2) onlayn korpuslaming qidiruv va qayta ishlash jarayoni uchun instrumentlar yaratish; 3) onlayn korpuslar va korpusga asoslangan instrumentlarni tanqidiy baholash; 4) korpusga asoslangan tahlilning metodologik masalalarini muhokama qilish; 5) lingvistika va kompyuter lingvistikasi, diaxron lingvistika, diskurs analiz, uslubiy analiz, tilga o'qitish kabi yondosh sohalardagi korpusga yo'naltirilgan dasturiy ta'minotlar yaratish va natijalarni o'rganish.

M.Kurdining e'tiroficha[30], korpus texnologiyasida lingvistlar uchun empirik tadqiqot olib borish imkoniyatining mavjudligi so'z turkumlarga ajratish teggeri, morfologik analizator, sintaktik parser kabi NLP sohasida yaratilgan ko'plab instrumentariylar yordamida amalga oshirish imkoniyati mavjud. Kompyuter lingvistikasi va korpus texnologiyasi o'zaro to'ldiruvchi va takomillashtiruvchi sohalar sifatida statistik bilimlarga ega bo'lish orqali yuqori natijalarga erishishda bir-birini taqazo etadi.

Tekstometriya sohasi ham korpus texnologiyasi bilan uzviy bog'liq. Mazkur yo'nalishda matnlar majmuasini metodologik jihatdan qayta tahlil qilish uchun lingvistik va matematik jihatdan sintagmatik hamda paradigmatik bog'lanishlar, qarama-qarshi xarakterlarni baholashda korpusdan olingan statistik ma'lumotlar natijasiga asoslanadi[31]. NLP sohasida matnlarni qayta ishlash va ma'lumotlarni turli jihatdan tahlil qiluvchi korpusni kompleks tahlil qilish (moslik darajasini aniqlash, klasterlash, leksik jadvallar yaratish, murakkab leksik qurilmalarni qidirish, turli o'lchovlarga ko'ra ichki korpuslarni ajratish) uchun samarali

[29] O'sha joyda. – B. 11
[30] Kurdi M. Z. Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax. – Great Britain, USA: Wiley-ISTE 2016. – P. 12
[31] https://txm.gitpages.huma-num.fr/textometrie/en/Introduction/

vositalardan biri TXM platformasi hisoblanadi[32]. Tekstometriya matn<-korpus->statistika uchligida ish ko'radi. Ushbu yo'nalishda 1970-yillarda Pierre Guiraud va Charlez Myuller tomonidan amalga oshirilgan ishlar sirasiga leksik statistikaga doir tadqiqotlarni keltirish o'rinli. Myuller matndagi lug'at hajmini aniqlash va ularni lingvistik tavsiflash borasida korpusga asoslanish zarurligini e'tirof etadi. Jen Paul Benzekri tekstometriyani ma'lumotlarni tahlil qilishning tadqiqot usuli sifatida tilshunoslik sohasiga olib kirgan[33]. Olim matnlar va so'zlarning sintetik va vizual xaritasini yaratish hamda ularning o'zaro bog'liq va qarama-qarshi jihatlarini umumlashtirish maqsadida korpus tahlilining amaliy ahamiyatiga e'tibor qaratgan.

Korpus texnologiyasi kompyuter lingvistikasida namunaga asoslangan (example besed machine translation) va statistikaga asoslangan neyro mashina tarjimasi rivojida katta rol o'ynadi. Shu kabi ishlar sirasiga *kontekstologik lug'at* asosiga qurilgan AMPAR mashina tarjimasi tizimini keltirish mumkin. Unga ko'ra, lug'at kontenti asos tildagi matn konkordansi va unga muqobil tarjima (parallel matnlar) variantlari asosida yaratiladi. Mazkur lug'at matn bilan bog'liq ravishda aniqlangan ko'p ma'noli yoki omonimlarning tarjima qilinayotgan matndagi lingvistik qurshovi bilan solishtiriladi. Mazkur lingvistik yondashuv asosida frazeologik birliklar, sintaktik strukturalar, morfologik kategoriyalar tarjima bosqichida tahlil qilinadi. Yana bir jihati parallel matnga asoslangan tahlil natijasida matnlarda aniqlangan konkordanslar ma'lumotlar bazasiga yig'iladi[34]".

Turkologiyada mazkur sohaga doir Kemal Oflazer[35] ishlarini qayd etish joiz. Ingliz tilidan turk tiliga tarjima qilishda statistik bilimlar bazasiga qurilgan mashina tarjimasi texnologiyasida ham korpus texnologiyasida erishilgan natijalarni ko'rish mumkin. Unda til juftliklaridagi grammatik qoidalardan tashqari

[32]Лаврентьев А.М., Соловьев Ф.Н., Чеповский А.М. Внедрение в TXM дополнительных инструментов автоматической обработки текста / PROCEEDINGS OF THE INTERNATIONAL CONFERENCE «CORPUS LINGUISTICS–2019», - С. 55.

[33] https://txm.gitpages.huma-num.fr/textometrie/en/Introduction/

[34] http://www.nop-dipo.ru/node Марчук Ю.Н. Типология текстов и машинный перевод.

[35] Ilknur Durgar El-Kahlout, Kemal Oflazer Initial Explorations in English to Turkish Statistical Machine Translation / Proceedings of the Workshop on Statistical Machine Translation. –New York City, 2006. – P. 7–14.

parallel matnlardan olingan bilimlar bazasi asos qilib olingan. Jumladan, 22500 gap fragmentlarida so'zlarning kontekstadagi holatiga ko'ra ingliz va turk tillarining grammatik jihatdam moslik darajasi aniqlangan.

Shuningdek, Ertuğrul Yilmaz, İlknur Durgar El-Kahlout, Burak Aydın, Zişan Sıla Özil, Coşkun Mermer kabi olimlarning ishlarida ham barqaror sintaktik modellar va iyerarxik frazaga asoslangan statistik mashina tarjimasi (turkcha-inglizcha) texnologiyasiga tayangan tarjima algoritmi ishlab chiqilgan. G.Ozbek[36]ishlarida parallel korpus texnologiyasi quyidagi bosqichlarda amalga oshirilgan: dastlab matnlardan olingan natija morfologik sathda ehtimollik nazariyasiga asoslanib, so'z shakllariga ajratiladi, so'ngra so'zshakllarning asosi aniqlanadi. Bu texnologiya dastlab Sabanji universiteti (Turkiya) professori K.Oflazer tomonidan ishlab chiqilgan. Tarjima xotirasida ingliz va turk tillaridagi turli janrga tegishli 22000dan ziyod gap juftliklari parallel korpusi yaratilgan. Shuningdek, alifbodagi nisbiy tafovut hisobga olinib, satr (string) tipidagi ma'lumotlarni umumiy formatga keltirish uchun matndagi barcha yozuvlar kichik shriftga o'tkazilib, milliy yunikod (unicod)dan standart yunikodga o'tkazilgan. Segmentlash jarayoni quyidagi juft fayllardan iborat: 1-sida so'zlarning asosi, 2-sida asos va affikslardan iborat ma'mumotlar bazasi. Keyingi fayl ushbu fayllardagi ma'lumotlarning qay tarzda bog'lanishiga javob beradi. So'zlarni ajratish uchun ushbu holatda ikki bosqichli Giza++[37]instrumentidan foydalanilgan. Ikki bosqichli texnologiyada dastlab bevosita juftlikdagi parallel korpuslar stemmizatsiya qilinadi (so'z shakllari asoslarga ajratiladi). So'ngra ichki moslashish jarayoni olib boriladi. Olimning fikricha, "$e_i \ldots e_n$" va "$f_i \ldots f_n$" parallel matnlardagi ikki shartdan biri mavjud bo'lsa, natija hisobga olinadi: agar e_i va f_i lug'atda mavjud bo'lsa, uning birligi sifatida yoki e_i va f_i lar ej (1 □□j □□n) yoki

[36]Gorkem O., Siddharth J. TU R K A L A T O R A Suite of Tools for Augmenting English-to-Turkish Statistical Machine Translation. 2006.

[37] Och Franz J. 2000. Giza++: Training of Statistical Translation Models. Available at http://www-i6.informatik.rwthaachen.de/~och/software / GIZA++.html.

fj (l □□j □□m) holati bilan lug'atda mavjud bo'lmaydi. Oxirgi jarayonda Giza++dan ikki fayl olinadi. Shu tarzda ma'lumotlar ustida amal bajariladi.

Matnni segmentlarga ajratish jarayoni gaplarni morfologik analizi uchun zarur sanaladi. Shu tarzda korpusdan olingan segment birliklarni stem (asos) va morfemalarga ajratish tarjima modelini voqelantirishda muayyan darajda rol o'ynaydi.

Mavzu yuzasidan savol va topshiriqlar:

1. Jahon korpus lingvistikasida qanday tadqiqotlar olib borilgan: umumiy va farqli jihatlari nimadan iborat?
2. Korpus boshqa yondosh sohalar bilan qanday aloqadorlikka ega, amaliy jihatlariga urg'u bering.
3. Turkologiyada korpus lingvistikasi bo'yicha olib borilgan eng so'nggi tadqiqotlardan (ilmiy maqolalardan) foydalanib, quyidagi jadvalni to'ldiring:

"B"- Bilar edim	**"B"- Bilib oldim**	**"B"-Bilishni xohlayman**

2-mavzu yuzasidan test

1. *Mega* va *milliy korpus* terminlariga izoh berishda Martin Weisser korpuslarning qaysi jihatiga e'tibor qaratgan?
   a) Qidiruv usullariga
   b) Qanday maqsadga yo'naltirilganligiga
   c) *Hajmiga
   d) Yaratilish texnologiyasiga
2. Hansard korpusi 1803-2005 yillarda og'zaki va yozma matnlardan yozib olingan qancha so'zni qamrab olgan?
   a) *1.6 mlrd
   b) 1.5 mlrd
   c) 1.9 mlrd
   d) 12.5 mlrd
3. Korpus lingvistikasida izlanish olib borayotgan tor doiradagi tadqiqotning metodologik usulini ham korpus deb nomlash mumkinligini e'tirof etgan olim ...
   a) Ch.Taylor
   b) *Ch.Mir
   c) N.Chomskiy
   d) K.Oflazer
4. Rus tilida yaratilgan birinchi korpus haqidagi tog'ri ma'lumotni toping.
   a) *1980-yilda Shvetsiyaning Upsala universitetida yaratilgan.
   b) 1980-yilda L.N.Zasorin rahbarligida yaratilgan.
   c) 1 mln so'zdan iborat bo'lgan.
   d) 1960-70- yillarda V.Zaxarov tomonidan Rossiyaning Moskva davlat universitetida yaratilgan.
5. Multimediali rus korpusi qanday matnlardan tuzilgan?
   a) *1930-2000-yillardagi kinofilmlar fragmentidan tuzilgan
   b) 1930-2000-yillarda yaratilgan badiiy va sahna asarlaridan foydalanilgan
   c) 2000-yillardagi film va sahna asarlarining matnlaridan tuzilgan
   d) 2001- yillardagi qo'shiqlar matnidan foydalanib tuzilgan
6. Rus tili milliy korpusi qanday manbalarni o'z ichiga oladi?
   a) *18-asr o'rtalaridan 21-asr boshigacha bo'lgan davriy manbalarni qamrab oladi
   b) 19-asr o'rtalaridan 21-asr o'rtalarigacha bo'lgan davriy manbalarni qamrab oladi
   c) 20-asr o'rtalaridan 21-asr boshigacha bo'lgan davriy manbalarni qamrab oladi
   d) 18-asr o'rtalaridan 20-asr boshigacha bo'lgan davriy manbalarni qamrab oladi

7. Kaedingning 1897-yilda korpus lingvistikasiga qo'shgan hissasi nimadan iborat edi ? U...
   a) *11 mlndan iborat nemischa so'zlardagi harflar ketma-ketligi va qo'llanish chastotasini o'rgangan.
   b) Bolalar tilini o'rganish uchun xorijiy til pedagogikasi bo'yicha ilmiy izlanish olib borgan
   c) Sintaktik razmetkaga asoslangan Penn korpusini yaratgan
   d) 1897- yilgacha korpus lingvistikasi qaysi bosqichlarni bosib o'tganligini tasniflab bergan
8. 1) 1960-yillar; 2) 1970-yillar; 3) 1980-yillar; 4) 1990-yillar.
   a-Brown, b-Upsala rus tilining korpusi (Shvetsiya) , c-BNC, d- LOB.
   Elektron korpuslarning xronologik yillar mosligi to'g'ri ko'rsatilgan qatorni toping.
   a) 1-b, 2-a, 3-c, 4-d .
   b) *1-a, 2-d, 3-b, 4-c.
   c) 1-b, 2-a, 3-d, 4-c.
   d) 1-a, 2-d, 3-c, 4-b.
9. V.Zaxarov tasnifiga ko'ra matnli fayllarni yartishdagi bosqichlar to'g'ri ko'rsatilgan javobni toping.
   1) reprezentatsiya; 2) grafematik analiz; 3) normallashtirish; 4) lemmatizatsiya; 5) tokenizatsiya; 6) stemmizatsiya.
   a) 1, 2, 4, 5
   b) 1, 2, 3, 6
   c) 1, 4, 5, 6
   d) *1, 2, 3, 4

10. Pushkinning "Yevgeniy Onegin" she'riy romanida qo'llanilgan rus tilidagi so'zlarning statistikasi kim tomonidan tayyorlangan?
    a) *A.A.Makarov
    b) A.S.Griboedev
    c) Chomskiy
    d) Mitkov

Glossariy				
1.	Mashina tarjimasi	машинный перевод	Machine translation	Bir tildan ikkinchi tilga og'zaki va yozma tarjima qilish jarayoni.
2.	ikki tilli parallel vositalar	двуязычное выравнива	bilingual alignment	bir matndan ikkinchi tilga

		ние	tools	o‘girilgan parallel korpus asosida bilingval lug‘at yaratishga mo‘ljallangan instrument.
3.	bilingual concordans	двуязычные конкордансы	bilingual concordance	ikki tilli parallel korpuslardan olingan leksik birliklar: termin, so‘z va iboralarning ikkinchi tildagi mos tarjimalari ro‘yxati.
4.	ikki tilli korpus	двуязычный корпус	bilingual corpus	ikki tilda tarjima qilingan parallel matnlarning elektron to‘plami.
5.	ikki tilli lug‘at	двуязычный словарь	bilingual dictionary	bir tildan ikkinchi bir tilga tarjima qilish uchun mo‘ljallangan lug‘at.
6.	chomskiy iyearxiyasi	иерархия Хомского	chomsky hierarchy	kontekstga bog‘liq bo‘lmay ichki guruhlarning muayyan tobelanish asosida yaratilgan paradigmatik munosabatlar yig‘indisi.
	qiyosiy korpus	сопоставимые корпуса	comparable corpora	bir nechta tildagi matnlardan olingan namunalarni solishtirish asosida yaratilgan korpus.
	kompyuter	компьютер	computational	lug‘atlarni