E-ISSN 2181-1466
ISSN 2181-6875

@buxdu_uz @buxdu1 @buxdu1 www.buxdu.uz

8/2023
Научный вестник Бухарского государственного университета
Scientific reports of Bukhara State University

BUXORO DAVLAT UNIVERSITETI

BUXORO DAVLAT UNIVERSITETI ILMIY AXBOROTI

8/2023

UDK: 81.432.1-3

KORPUS LINGVISTIKASI TARAQQIYOTI VA O'ZBEK TILSHUNOSLIGIDA KORPUS TAHLIL ASOSLARI

Raximov Mubin,
Buxoro davlat universiteti o'qituvchisi,
mubinrahimovbuba@gmail.com

Annotatsiya. *Mazkur ilmiy maqola tilshunoslikning bir bo'lagi bo'lmish korpus lingvistikasining taraqqiyoti va bugungi kundagi muammolarini o'rganishga bag'ishlangan. Maqolada ilk konkordanserlarning yaratilishidan tortib eng so'nggi avlod konkordanser dasturlari ilmiy tahlil qilingan. Har bir avlod dastur konkordanserlarning funksiyasi va ustunlik hamda kamchiliklari taqqoslanib, o'zbek tilidagi dasturlar yaratilishi bo'yicha ilmiy izlanishlar ko'rsatib otilgan.*

Kalit so'zlar: *korpus lingvistikasi, konkordanserlar, lingvistika, korpus tahlil*

РАЗВИТИЕ КОРПУСНОЙ ЛИНГВИСТИКИ И ОСНОВЫ КОРПУСНОГО АНАЛИЗА В УЗБЕКСКОМ ЯЗЫКОЗНАНИИ

Аннотация. *Данная научная статья посвящена изучению развития и современных проблем корпусной лингвистики, которая является разделом языкознания. В статье проводится научный анализ от создания первых конкордансеров до последнего поколения программ-конкордансеров. Сравниваются функции, преимущества и недостатки каждого поколения программ-конкордансеров, а также показаны научные исследования по созданию программ на узбекском языке.*

Ключевые слова: *корпусная лингвистика, конкордансеры, лингвистика, корпусный анализ.*

DEVELOPMENT OF CORPUS LINGUISTICS AND THE BASIS OF CORPUS ANALYSIS IN UZBEK LINGUISTICS

Abstract. *This scientific article is devoted to the study of the development and current problems of corpus linguistics, which is a branch of linguistics. In the article, from the creation of the first concordancers to the latest generation of concordancer programs, there is a scientific analysis. The functions and advantages and disadvantages of each generation of program concordancers are compared, and scientific research on the creation of programs in the Uzbek language is shown.*

Key words: *corpus linguistics, concordancers, linguistics, corpus analysis*

Kirish. Tilshunoslikning taraqqiy etishi tahlillar hamda statistik ko'rsatkichlar o'rtasida uyg'unlik kasb etib kelmoqda. Har bir tahlil struktural yoki morfologik bo'lishidan qat'iy nazar, lisoniy tadqiqotlarda tillararo qiyoslash asosida amalga oshirilgan izlanishlarning natijadorlik ko'rsatkichi yuqori bo'ladi.

Korpus tilshunosligi atamasining XIX asr 60 – yillarida fan sifatida yuzaga kelishi til me'yorlarining nafaqat statistik, balki struktural va morfologik ko'rsatgichlar tahlilini ochib berishda asosiy me'zon bo'lib xizmat qildi. Matnlar korpus tahlili XIII asrda vujudga kelgan bo'lib, bunga yorqin misol xristian Injiliga A. Kruden tomonidan tadqiq qilingan Archeotypical corpus (arxeotipik korpus) bo'lib, yig'ilgan matnlar word indexing (so'zni indekslash) yoki concordancing (so'zlarning o'zaro munosabati va bog'lanishi) orqali tartib bilan joylashtirilgan. [21] Bu esa o'z navbatida korpus lingvistika taraqqiyotining *I bosqichi elektronlashtirishgacha bo'lgan davr*ni boshlab berdi. Ushbu davrning o'ziga xosligi shundaki, bunda diniy asarlarga qo'lyozmada konkordanslar (korpus matnida qatnashgan alfavit tartibiga keltirilgan so'zlar).

Asosiy qism. A. Kruden Injilda ishlatilgan lug'at yoki yo'naltiruvchi so'zlarni alfavit tartibidagi ko'rinishi va shuningdek ularning joylashgan o'rnini belgilaydigan birliklarni "konkordanalar" deb ataydi.[6] Bundan ko'rinib turibdiki, konkordansli linguistik tahlillar zamonaviy korpus tahlillardan farqli o'laroq faqat so'zlarning konteksdagi alfavitli tartibi hamda ishlatilish o'rnini (index) belgilab bergan. 1890-yilda J. Strong *(Strong's Exhaustive Concordance of the Bible)* "Injilning to'liq konkordansi (muvofiqligi)" kitobini nashr etadi, unda Eski Ahddagi 8 674 ta ildizi ibroniycha so'z va Yangi Ahddagi yunon tilidagi 5 624 ta so'z uchun etimologik ma'lumot keltiriladi. Har bir so'z uchun ishlatilish miqdori (chastotasi) va foydalanilgan o'rni haqida ma'lumot beriladi. [16]

Keyinchalik konkordansli tahlillar badiiy asarlarlarga nisbatan amalga oshirildi. A. Bekket tomonidan Uilliam Shekspir asarlarining barcha nashrlariga yaratilgan konkordanslar" asarida muallif nafaqat ma'lum bir so'zning joylashuv o'rni va miqdori haqidagi ma'lumot balki, mazkur so'z qatnashgan parchalar ham keltirib o'tiladi. Bunga misol qilib muallif tomonidan tahlil qilingan "*dream*" (tush) so'zi ko'rib chiqilgan. Muallif "*dream*" so'zining birlik va ko'plik formasini ishlatgan holda besh qatordan iborat parcha keltiradi. Tahlil qilingan so'z uning barcha so'z shakllarini o'z ichiga oladi *(dream, dreams*). [2; 470]

Shuningdek Shekispir asarlarining konkordanslari M. Kouden -Klark (1847) va S. Oyskot (1790) tomonidan ham tadqiq qilindi. Ularning tadqiqotlarida so'z va barcha so'z shakllari konkordanslar sifatida tadqiq qilindi. Konkordanslarni an'anaviy (manual shaklda) tadqiq qilish an'analari 1995 yilgacha davom etdi.

Konkordanslarning taraqqiyotidagi tub burilish Keyword out of context (KWOC) kontekstdagi kalit so'zlar yoki Keyword in title (KWIT) sarlavhadagi kalit so'zlarni tadqiq etish metodikasini ishlab chiqish bo'ldi. Bu tizim 1856 yilda A. Krestadoro tomonidan Manchester davlat kutubxonasidagi kataloglarlarni tizimlashtirishda ishlatilgan. 1958 yilda X. P. Lun tomonidan ushbu metod qayta ishlanib, keywords in the context (KWIC) -matndagi kalit so'zlar -tizimi kompyuter texnologiyasi orqali qayta yaratildi. Bu tizimda kalit so'z markazda joylashib uning ikki tomonida konkordanslar tizimi joylashtirildi.[11;149-171] KWIC formati alifbo tartibida so'z birikmalarining ro'yxatini tuzishga imkon beradi, shuningdek har bir ishlatilgan so'zning chastotasini belgilaydi. Keyinchalik bu atama elektron konkordans deb atala boshlandi. Rohib R. Busa tomonidan Foma Akvinskiy ishlariga yaratilgan, 10,6 million so'zlarni o'z ichiga olgan elektron konkordans -Index Tomisticus mashina ishlanmasi ko'rinishidagi birinchi tadqiqot bo'ldi. [4; 83-90] Tadqiqot ishlari 4 yil 1962-1966 yillarda olib borildi. Konkordanslar bilan ishlashni qulay hamda sodda bo'lishi uchun R. Busa kalit so'zlar o'rnida faqatgina lemmalar yoki barcha so'z shakllarini mujassamlashtirgan bosh so'zlarni tanladi. Buning uchun olim ikki bosqichda matnning lemmalizatsiyasini o'tkazadi: barcha fleksiyali so'z shakllarini bir lemma ostida birlashtirish va har bir lemma hamda so'z shakllari uchun uning nutq bo'lagiga mos keldadigan kodini biriktirish. Lemmalizatsiya jarayoni olimning o'zi va o'nta rohib tomonidan tuzilgan Lotin mashinali lug'ati (Lexicon Electronicum Latinum) asosida olib borildi. Elektron lug'at o'zida lemmalar jadvalini mujassamlashtirgan bo'lib, kompyuter ular orqali lemmetizatsiyalashni amalga oshirdi. Elektron lug'at asosida olib borilgan tadqiqot metodi keyinchalik matnlarni elektron qayta ishlash tamoyilini belgilab berdi.

Korpusning elekrton davrigacha bo'lgan eng oxirgi tadqiqot R. Kverk tomonidan London universitetida ishlab chiqilgan og 'zaki va yozma nutq qo'shma korpusi "Ingliz tilining amaliy ishlatilish sharhi" The Survey of English Usage, SEU bo'ldi.[14] Kverk yig'ilgan tadqiqot materiallarini "birlamchi manbalar" yoki "matnlar" deb belgiladi. Mazkur korpus o'z davri uchun yetarlicha sistemalashtirilgan va strukturalashtirilgan og'zaki va yozma nutq manbalarini mujassamlashtirgan korpus bo'lib xizmat qildi. Bu korpus rasmiy va norasmiy nutqqa oid 200 ga matn namunalarini mujassamlashtirgan 5000 so'zdan iborat qilib tuzildi.

Shunday qilib korpus lingvistikasining elektron davargacha bo'lgan tadqiqotlarda konkordanslarning yaratilishi hamda ular lug'at yoki ko'rsatgichlar sifatida ko'rilishi korpusning yuzaga kelishiga asos yaratdi. Mazkur konkordanslarni korpus sifatida qabul qilinmasligining asosiy kamchiliklari bu ularda diniy kitoblar, eski davr badiiy adabiyotlardagi matnlarni tadqiq qilinishi bo'ldi. Binobarin, zamonaviy tilshunoslik nuqtayi nazaridan bu konkordanslar korpus emas, balki arxiv sifatida qabul qilinadi. Tadqiqotlarning yana bir kamchiligi korpuslarning yig'ishning ma'lum bir tamoyili va konkordanslarni tuzishning yagona tartibi mavjud emas edi. Shunday bo'lsa-da, ilk konkordanslarning yaratilishi korpus linguistikasining taraqqiyotiga muhim qadam hisoblandi, chunki aynan qidirilayotgan so'zlarning ko'rsatib o'tilishi hamda ularning matndagi o'rnini belgilab berilishi majburiy aspekt sifatida qarab kelindi.

Korpus lingvistikasining *elektronlashtirilgan II – bosqichi* 1960-yilardan hozirgi kunga qadar davom etib kelmoqda. Olim S. Yoxansonning fikricha korpus lingvistikasi 1970-yillarga kelib soha sifatida rivojlana boshladi. To'g'ri 1960-yillarda R. Busa tadqiqotlari korpus linguistikasining boshlanishi bo'lsa-da, matn korpuslarini tadqiq qilish ilk laboratoriyalari va kompyuter markazlari 1970-yillarda vujudga kela boshladi, unda tilshunos olimlar bilan bir qatorda kompyuter mutaxassislari ham faoliyat ko'rsata boshladi.[9;33-35] Italiya, Angliya, AQSh, Kanada, Fransiya, Germaniya, Norvegiya, Shvetsiyada matn korpusini to'plashga, qayta ishlashga va saqlashga lingvistik markazlar tashkil etildi. 1970-yillarning o'rtalariga kelib elektron korpuslarni saqlay oladigan va tarqata oladigan bazalar: Oksford mashinada o'qiladigan matnlar arxivi OTA (Oxford text archives) (1976) va Zamonaviy ingliz tilining elektron matnlari arxivi ICAME (International Computer Archive of Modern English) (1977) yaratildi.

XX asrning 60-yillarida AQSh Bravn Universiteti (The Brown University)da Bravn korpusining yaratilishi elektron korpusning ilk bosqichini boshlab berdi. G. Kuchera va N. Frensis boshchilik qilgan mazkur loyihada amerika yozma nutqining 15 janridagi 500 matndan iborat bir millionta so‘z o‘rin olgan.[24] Ushbu korpusning o‘ziga xosligi, bu undagi matnlar perfokartaga kiritilgan bo‘lib, matn joylashuv o‘rni, nomlanishi hamda matndagi qatorlar soni haqidagi ma’lumotlarni o‘zida mujassamlashtirdi.

1968-yilga kelib olim F. Begli ilk marotaba matn korpusidagi barcha ma’lumotlarni ifodalovchi *metadata* terminini kiritdi. [13;189-215] O‘tgan asrning 60-yillarida KWIC asosida ishlaydigan ilk konkordanser-dasturlar: “korpus hisobi va konkordanslarini yaratuvchi atlas” (COCOA, COunt and COncordance Generation Atlas) (1967) va «Kollokatsiya» (CLOC, CoLOCation) (1978).[17;2] paydo bo‘ldi. Bu ikki dasturning yaratilishida avtomatik qayta ishlanish jarayoni ma’lumotlarini mujassamlashtirgan *kod* yoki *teg*larni matnga qo‘lbola biriktirish orqali qo‘llab-quvvatlandi.[1;154] To‘liq avtomatlashtirish haqida esa faqatgina 1971-yilda B. Grin va J. Rubbin tomonidan yozilgan matnlarini avtomatik belgilash dasturi TAGGIT yuzaga kelgandan so‘ng gapirila boshlandi. TAGGIT dasturi matnda ajralib turadigan muhim va rasmiy so‘zlarni, tinish belgilari hamda alohida morfemalarni 86 ta *teg* orqali belgilab berdi. Dastur "omonimiyani olib tashlamadi" va natijada korpusdagi so‘zlarning 23 foizi bir vaqtning o'zida bir nechta teglar bilan belgilandi.[12;312] Gap bo‘laklarini belgilovchi Braun korpusini 1979 yilga kelib tuzatishlar va approbatsiyadan o‘tkazilishi yakunlandi. B. Grin va J. Rubinlar Braun korpusi asosida yaratilgan TAGGIT analizator dasturi haqidagi barcha ma’lumotlarni basharti uni qayta ishlash va mukammalashtirib borish uchun ochiqlantirishdi. [9;46] Birinchi avlod konkordanser-dasturlar COCOA va CLOC har qaysi alohida kompyuterlar uchun alohida topshiriq sifatida yaratilib borilishi sabab, ularni har doim qaytadan dasturlash lozimligini taqozo etdi. Bu esa birinchi avlod konkordanser-dasturlarining asosiy kamchiligi edi va bu holat ikkinchi avlod konkordarserlarini yaratilishini taqozo qildi. Xulosa qilib aytganda, birinchi avlod konkordanser dasturlari va analizatorlari yaratilishi olimlar ta’biri bilan aytganda 1970 yillar so‘ngiga kelib korpus lingvistikasiga fan sifatida asos solinishiga olib keldi.[18;12]

1980-yillarda TAGGIT analizatorini yanada mukammalshtirish ishlari davom ettirildi va Lankaster universitetida bir qator olimlar grammatist J. Lich hamda dasturchi R. Garsayd boshchiligida CLAWS (the Constituent Likelihood Automatic Word-tagging System - O'xshashlikka asoslangan avtomatik komponentlarni etiketlash tizimi) nomli yangilangan morfologik analizatorni tadbiq etishdi. [12]

“Braun korpusi” boshqa korpuslarni yaratish uchun asos bo’lib, unda yozma nutqning miqdoriy va spektoral uslubi ustunlik qilib keldi. 1970-yillar o‘rtalariga kelib mazkur korpusning chop etilishi avval Buyuk Britaniya keyinchalik boshqa davlatlarda turli korpuslarning yaratilishiga zamin bo’ldi. Misol qilib aytadigan bo’lsak, 1976-yilda Lankster, Oslo va Bergen universitetlarining qo’shma korpusi (The Lancaster-Oslo-Bergen corpus (LOB) yaratildi.[25] Ushbu korpusga 1990-yilga kelib ko’plab analoglar yaratila boshlandi: Avstraliya korpusi-Australian Corpus of English, ACE (1986); Yangizelland ingliz yozma nutqining Vellington korpusi, The Wellington Written English, WWE (1986), Frayburg va Braun universitetlarining amerika ingliz tilisi korpusi The Freiburg-Brown Corpus, FROWN (1991–1992); Frayburg, London, Oslo va Bergen universitetlari ingliz tili korpusi-The Freiburg London-Oslo / Bergen corpus, F-LOB, (1991–1992); Hindiston ingliz tilisi yozma nutqining Kolxapur korpusi The Kolhapur corpus Indian English (1978). [10] Ushbu korpuslar “Braun korpuslari oilasi” nomini olgan bo’lib, ularning eng asosiy farqli jihati ulardagi yozma nutq matnlari amerika, britaniya, avstraliya yoki hind ingliz tili variantida kiritilishi bo’ldi. [20, 383-457] *Og’zaki nutq korpuslari* yozma nutq korpusiga nisbatan ancha kechroq 1975-1990 yillarda Ya. Svartvik, R. Kverk, S. Grinbaum va K. Xolland boshchiligida London -Lund korpusini yaratish bilan yuzaga kela boshladi. Mazkur korpus ikki: SEU (1959–1989) va og‘zaki nutq korpusi (SSE, 1975) asosida yaratildi. Unda jami 100 ta transkriptlagan har biri 5000 so‘zdan iborat monologik va diologik nutqdan tarkib topgan bo’lib, diologik nutq so‘zlashuv uslubidagi do‘stlar va hamkasblar muloqotini, monologik nutq esa to‘g‘ridan-to‘g‘ri hikoya hamda izohlarni qamrab olgan. [20;383-457] 1992-1994-yillarda SEC corpusi asosida yangi korpus MARSEC (Machine-readable spoken English corpus) ham yaratilib bunda morfologik tahllillardan keng foydalaniladi.

Bravn korpusi yaratilishi bilan bilan bir qatorda yan bir tushuncha “referent korpus” tushunchasi yuzaga keldi. Referent korpuslar asosan matnning turli birliklari chastotalarini aniqlash uchun ishlatila boshlandi. [1;137]

Aynan 90-yillarga kelib og’zaki nutqning turli birliklarini tahlil qiladigan korpuslsr yaratila boshlandi. Og‘zaki nutqni aniqlay olish va sintezlash maqsadida AQShda (Defense Advanced Research Projects Agency, DARPA), talaffuzda raqamli ketma-ketlikni aniqlasy oladigan TI-DIGITS (1984 Texas Instituti AQSh), sheva nutqlarni aniqlashda ishlatilgan TIMIT korpusi (1990 yil AQSh), Resurslarni boshqarish korpusi RMC (1988 yil AQSh) shular jumlasidandir.

1990-yillarda AQSh da harbiy buyurtmalar asosida ham turli korpuslarni yaratish ishlari jadallashdi. Bunga misol qilib, ATIS (Air Travel Information Service Corpus) korpusini olish mimkin. Mazkur korpus muntazam og‘zaki nutqni tanib olish va sintezlash uchun ishlatildi.[19] Mazkur harbiy korpuslar quyidagi yangi terminlarning yaratilishiga olib keldi: tokenlash (uzluksiz gapni alohida so‘zlarga ajratish), segmentatsiya (uzluksiz gapni gap va sintagmalarga ajratish), parser (sintaktik analizator), normallashtirish (fonetik qisqartirish, turli individual xususiyatlar bilan talaffuz qilinadigan so‘zlovchi so'zlarning normasi) va h-zo.

Umumiy qilib aytganda o‘z taraqqiyotining birinchi davrida korpus atamasi to’liq shakllandi. Korpusda yangi atamalar vujudga keldi: “korpus lingvistikasi”. “belgilash”, “metabelgilash”, “konkordanser”, “morfologik analizator”, “tokenlar”, “tokenlash” “sigmentatsiya”, “normallashtirish”, “vaqtga tenglashtirish” (time alignment) shular jumlasidandir.

O‘gaki nutqning rivojlanishi bilan 1990-yillarda Lankster Universiteti olimlari tomonidan quyidagi belgilash dasturlari ishlab chiqildi: anaforik-referentlik bog‘liqlikni belgilash (1992), prosodik belgilash (1993), semantik belgilash (1993), (2004), badiiy stilistik (1996 и 2004), programmali belgilash (2003) va so‘zlovchining xatolarini belgilsh (1999, 2003).[12;78,83] T. Makineri va A. Xardilarning ta’kidlashicha 1990-yillar konkordanserlar taraqqiyotining *keying davr ikkinchi avlodi* yaratilishini boshlab berdi. Ikkinchi avlod konkordanserlarining xususiyatlari bu ularning IBM operatsion tizimida ishlashi va ular barcha personal kompyuterlarida qo‘llanila olish imkoni borligi edi. Micro-OCP (1988), Longman Mini-Concordancer (1989), Kaye concordancer (1990) - ikkinchi avlod konkordanserlari ham KWIC metodikasi asosida ishlar edi. Ularning asosiy funksiyasi har ikkala tomonidan kontekstual so‘zlar bilan birikkan konkordarserlari alfavitli tartibini yaratish, korpusdagi so‘zlar ro’yxatini tuzish, elementar tavsiflovchi statistik ma'lumotlar, masalan, so‘zlardan foydalanish soni, so‘zlar soni va so‘zlardan foydalanish nisbati (type-token ratio) aniqlashdan iborat edi. Bularning asosiy kamchiligi esa belgilarni aniqlashtirishning yagona formati hamda standartini mavjud bo’lmagani edi.[12;40]

Yuqoridagi muammolarni hal qilish uchun yagona kodlash tizimini yaratish zarurati paydo bo‘ldi. 1991-yilda “Unikord konsotsium” notijorat kompaniyasi tomonidan ASCII (American Standard Code for Information Interchange) uchun “Unicode” belgilarni kodlash standartini ishlab chiqdi. Ushbu tizim barcha dunyo yozuv tillari hamda chop etilmaydigan belgilarini (transkripsiya, matematik formulalar va boshqalar) kodlash uchun qo’llanildi. Bugungi kunda UFT-8 unikordning keng tarqalgan shaklidir.[1;37,38]

Yevropada ham kodlashtirish tizimini ishlab chiqishga alohida e’tibor qaratildi. 1998 – yilda evroppalik mutaxassislar tomonidan matnlarni to’plash belgilash bo’yicha korpus Corpus Encoding Standard (CES) ishlab chiqildi. Mazkur korpuslar asosida ma‘lum bir standartga keltirilgan matnlar yoki ularni kodlash tizimi joy oldi.

1993 -yilda G. Lich tomonidan taklif qilingan “metamatn” tushunchasi, ya‘ni to‘laligicha ekstralinguistik ma‘lumotlarni o‘z ichiga olgan matn haqidagi matn, korpusning keyingi avlod taraqqiyotini belgilab berdi.[12;275-281]

Korpus lingvistikasi taraqqiyotining II davri 1990 -yillarning oxiri hamda 2000-yillarning boshlarida *uchinchi avlod konkordanserlari* (WordSmith 0.4 (1996), MonoConc (2000), AntConc (2005)) yaratildi va tatbiq etila boshlandi. Uchunchi avlod konkordanserlarining ustunlik tomonlari ko’p bo’lib, ular istalgan yozuv shaklidagi katta hajmdagi matnlarni qayta ishlashi, murakkab statistik tahlillarni olib borishi, katta tezlikda kalit so‘zlar va konkordanslarning ro‘yxatini ishlab chiqishi, chastotali hamda kolokatsion tahlillar qila olishi mumkin edi.[12;35]

Shunday qilib 1990 -yillarga kelib juda katta hajmdagi matnlarni yig‘ish va tahlil qilish imkoniyatlari yuzaga keldi. Shuningdek og‘zaki korpuslarni prosodik, fonetik, morfologik, leksik, sintaktik va diskursiv darajada tahlil qilish imkoniyati paydo bo‘ldi. Bu esa konkordanslarni avtomatik tarzda qayta ishlovchi bir necha dasturlarning yuzaga kelishini ta’minladi. G. Kennedi, P. Beyker, A. Xardi, T. Makineri o‘z asarlarida ushbu davrni megakorpuslar davri deb atashdi, chunki aynan shu davrda The Longman Corpus Network (1991), The Bank of English, BoE (1993), The British National Corpus, BNC (1994), The American National Corpus, ANC (2008) kabi tarkibida 100000000 (yuz million) ortiq so‘zga ega korpuslar yaratildi. Korpuslarning o‘ziga xosligi shunda ediki, ularda 75 % yozma nutq matni 25% og’zaki nutq matnidan tashkil topgani va istalgan turdagi konkordanserlarning tahlilini qilish imkonini mavjud edi.

Umumiy qilib aytganda, ikkinchi avlod korpuslari 100000000 (yuz million) so‘zdan kam bo‘lmagan baza, deyarli barcha janr hamda uslublardagi og‘zaki hamda yozma nutq materiallarining tahlil qilinishi va fonologik tahlil qilish imkoniyatlarining mavjudligi ularni keyingi taraqqiyotiga zamin yaratdi.

*Uchinchi avlod korpuslari*ning yaratilishida 2010-yillarda BNCweb (2009), CQPweb (2012), SketchEngine (2013), Wmatrix (2013) kabi *to’rtinchi avlod konkordanserlari*ning ixtiro qilinishi sabab

bo'ldi. Ushbu yangi avlod korpuslari "gigakorpuslar" deb ham atala boshlandi, chunki ularda juda katta hajmdagi matnlarni (ya'ni bir milliardga yaqin so'zlardan tashkil topgan) tahlil qilish imkonini beruvchi konkordanserlar shakllantirila boshlandi. M. Devis to'rtinchi avlod konkordanserlarini gibrid korpuslar deb ataydi, bunga sabab qilib esa ularning interfeysi morfemali, leksik, sintaktik va frazeologik daraja jihatidan chastotali tahlillar va korpus yaratishda maydon bo'lib xizmat qildi.[7;11-31] Bu davrda (COCA, Google Books Ngram) korpusining yaratilishi hamda unda bugungi kunga kelib milliarddan ko'p so'z hajmiga ega bazaning shakllantirilishi korpus lingvistikasi taraqqiyotining modernizatsiyalashgan bosqichini boshlab berdi. Katta hajmli korpuslar uch, to'rt va undan ko'p kollokatsialarni (so'z birikmalarini) chastotali tahlil qilish imkoniyatlarini bera boshladi. Mazkur turdagi kollokatsiyalarni D. Bayber [3] va K. Haylend [8;4-21] "leksik to'plamlar" deb atashdi, chunki kollokatsiyalardagi bitta so'z o'zgaruvchan bo'ladi. Misol qilib aytadigan bo'lsak besh so'zdan iborat *in the beginning of the, in the end of the, in the form of the* kollokatsiyalarda uchinchi so'z o'zgarib turadi. Keyinchalik bunday turdagi so'z birikmalari *n-gramm*lar deb atala boshlandi.[15;39-49]

2008 -yilda zamonaviy amerika ingliz tilisi korpusi (The Corpus of Contemporary American English (COCA) yaratildi. Mazkur korpus juda katta hajmga ega bo'lib, dinamik tarzda har yili 20-30 million so'zlar bilan boyitilib boriladi. Korpusning asosiy banki 1800-yildan to hozirgi kunga qadar nashr qilingan, e'lon qilingan, so'zlangan og'zaki va yozma nutq matnlaridan tashkil topgan. Bugungi kunga kelib ushbu korpus AQSh va dunyo miqyosida amerika ingliz nutqi tahlilini olib boruvchi eng asosiy korpuslardan biriga aylandi.

Yana bir mukammal korpus 2009-yilda yaratilgan raqamlashtirilgan kitob matnlari korpusi Google Books Ngram Viewer 1500-yildan to 2008-yilga qadar nashr qilingan elektron kitoblarning matnini o'z ichiga oldi. Mazkur korpusning bugungi kundagi bazasi 200 milliardan ko'p so'zga ega. Ingliz tilidan tashqari yana 6 ta tilda ham ushbu korpus ba'zalari yaratildi: nemis, fransuz, ispan, rus, italyan va yahudiy tillarda. [22]

Ingliz tilining hududlar kesimidagi veb-matnlarni tahlil qilish uchun 2013 – yilda Global Web-based of English (GloWbE) korpusi yaratildi. Mazkur korpusda barcha hududiy ingliz tili veb sahifa matnlarini o'z ichiga olgan 1,9 milliard so'z ba'zasi mavjud.[23]

Yuqoridagiga o'xshagan korpus News on the Web (NOW) 2016 – yilda yaratilgan bo'lib, bugungi kunda 5,7 milliarddan ko'p so'zga ega ba'zasi mavjud. данный момент превышает 5,7 миллиарда словоупотреблений. Korpus 2012 yildan to bugungi kunga qadar ingliz tilidagi matnlarini o'z ichiga oladi. Har kuni mazkur korpusga 4-5 million so'z qo'shiladi.[23]

Korpus lingvistikasiga qiziqish o'zbek tilshunoslari tomonidan yangi soha sifatida XXI asrga kelib mavjud korpuslarni keng qo'llash, o'zbek tili korpuslarini yaratish masalalarini o'rganishda namoyon bo'lib kelmoqda. Jumladan, J. Jumaboyeva ingliz tili korpuslaridan BNC tarkibidagi semantik gradatsiyalarni topish, G. Ergasheva BNC korpusi orqali genderga oid konseptual metaforalarni tadqiq qilish, G. Sobirova BNC korpusini tillarni o'qitishda qo'llashga e'tibor qaratgan. Sh. Hamroyeva mualliflik korpusini yaratish muammosiga yechim izladi.[5] Bundan tashqari N. B. Ataboyev COCA korpusi misolida Ingliz tilining diaxronik korpuslarining funksional xususiyatlarini o'rganib chiqdi.

Xulosa. Xulosa qilib aytilganda, ingliz tili korpusi taraqqiyotida ko'plab tadqiqot va izlanishlarni olib borgan bo'lib, uchinchi avlod konkordanser dasturlari og'zaki va yozma nutqdagi matnlarning mukammal megakorpuslarini yaratilishga hamda korpus lingvistikasining mukammal darajada taraqqiy etishiga olib keldi. Bunday korpuslar o'zida milliardlab so'zlar bazasidan iborat bo'lgan ingliz tilining turli janr, uslub hamda dialektual xususiyatlarini tahlil qilish imkonini berdi. Ingliz tili korpuslarining yagona kamchiligi bu ularda hissiyot tuyg'usi bilan ifodalangan nutq matning tahlil qilish masalalarini tadqiqot hamda izlanish olib borish ehtiyoji mavjudligidir.

O'zbek tilshunosligida korpus lingvistikasi yangi kam o'rganilgan soha bo'lib, avvalo o'zbek tili nutqiy matnlarning lug'at ba'zalarini shakllantirish, saqlash hamda tahlil qilish masalasidek qator vazifalar turibdi. Bugungi kunda asosan mavjud boshqa til (ingliz, fransuz, rus va hokazo) korpuslarini o'rganib ilmiy izlanishlarda tatbiq etish, bu korpuslar asosida o'zbek tilidagi matnlarni ham tahlil qilish va qiyoslash ishlari soha vakillarining ilmiy izlanishlarida o'z o'rnini topgan bo'lsa-da, XXI asrda mamlakatimizda barcha sohalarni dasturlash hamda raqamlashtirish dolzarbligicha qolmoqda.

ADABIYOTLAR:

1. Baker P., Hardie A., McEnery T. Glossary of Corpus Linguistics. Edinburgh University Press, 2006. 192 p

2. Becket A. A Concordance to Shakespear: suited to all the editions. 1787. P.470.

3. Biber D. University Language: A Corpus-based Study of Spoken and Written Registers. Amsterdam : John Benjamins, 2006. 261 p.

4. Busa R. The Annals of Humanities Computing: The Index Tomisticus //Computer and Humanities. 1980. Vol. 14. P. 83-90.

5. N. B. Ataboyev. Ingliz tilining diaxronik korpuslarining funksional xususiyatlari (COCA misolida) /monogragiya. Muhammad poligraf. 2023. b.5.

6. Cruden A. A Complete Concordance to the Holy Scriptures of Old and New Testament. London. (1737).

7. Davies M. Corpora: an introduction // The Cambridge handbook of Corpus Linguistics / ed. by D. Biber, R. Reppen. Cambridge University Press, 2015. P. 11–31.

8. Hyland K. As it can be seen: Lexical bundles and disciplinary variation // English for Specific Purposes. 2008. Vol. 27. P. 4–21.

9. Johansson S. Some aspects of the development of corpus linguistics in the 1970-s and 1980-s//Corpus Linguistics: An International handbook/ed.by A. Ludeling, M. Kyto.2008. P. 33-35.

10. Kennedy G. An Introduction to Corpus linguistics. Addison Wesley Longman limited, 1998. 315 p.

11. Koriyenski C., Newell A. F. The Indexing: the problem of the significance//Computers and the writing. State of the Art led by P.O.Holt [et.al] 1992. P. 149-171.

12. McEnery T., Hardie A. Corpus Linguistics: Method, theory and practice. Cambridge university press, 2012. 312 p.

13. Nguen T.H., Nunavath V., Prinz A. Big Data Metadata Management in small Grids // Big Data and Internet of Things: A Roadmap for Smart Environments. 2014. P. 189–215.

14. Quirk R. A Grammar of Contemporary English. 1972. 1120 p.

15. Rayson P. Computational tools and methods for corpus compilation and analysis // The Cambridge handbook of English corpus linguistics / ed. by D. Biber, R. Reppen. Cambridge university press, 2015. P. 32–49.

16. Strong J. (1890) Strong's Exhaustive Concordance of the Bible. The Methodists Book Concern.

17. Stubbs J. Notes on the History of Corpus Linguistics and Empirical Semantics // Collocations and Idioms / eds by M. Nenonen, S. Niemi. Joensuu: Joensuun Yliopisto, 2007. P. 2.

18. Svartvik J. Corpus linguistics 25+ years // Corpus Linguistics 25 Years On / ed. by R. Faccinetti. 2007.

19. Tur G. Spoken Language Understanding: Systems for Extracting Semantic Information from Speech / ed. by G. Tur, R. De Mori. 2011. 470 p.

20. Xiao R. Well-known and influential corpora // Corpus Linguistics: An International Handbook / ed. by A. Ludeling, M. Kyto. 2008. P. 383–457.

21. Солнышкина М.И., Гатиятулина Г.М. История развития корпусной лингвистики (на примере англоязычных корпусов)//Вестник Томского государственного университета № 63, - Томск, 2020. – С. 132. https://cyberleninka.ru/article/n/istoriya-razvitiya-korpusnoy-lingvistiki-na-primere-angloyazychnyh-korpusov/viewer [дата оброшения: 25.08.2023]

22. Google Books. URL: https://googlebooks.byu.edu/ [murojaat qilish sanasi: 08.08.2023].

23. GloWbE. URL: https://corpus.byu.edu/glowbe/ [murojaat sanasi: 08.08.2023].

24. The Brown Corpus. URL: https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpusling/content/corpora /list/private/brown/brown.html [murojaat sanasi: 30.08.2023]

25. The LOB Corpus. URL: http://www.helsinki.fi/varieng/CoRD/corpora/LOB/index.html (дата обращения: 31.08.2023)